UNIVERSIDADE FEDERAL DE GOIÁS
ESCOLA DE ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO

VIVIANE MARGARIDA GOMES

Controle Inteligente de Tempo Livre em Tutoria Multissessão

Goiânia
2009

VIVIANE MARGARIDA GOMES

Controle Inteligente de Tempo Livre em Tutoria Multissessão

Dissertação apresentada ao Programa de Pós–Graduação do Escola de Engenharia Elétrica e de Computação da Universidade Federal de Goiás, como requisito parcial para obtenção do título de Mestre no Programa de Mestrado em Engenharia Elétrica e de Computação.

Área de concentração: Engenharia de Computação

Orientador: Prof. Weber Martins, Ph.D.

Co–Orientador: Prof. Lauro Eugênio Guimarães Nalini, Dr.

Goiânia
2009

VIVIANE MARGARIDA GOMES

Controle Inteligente de Tempo Livre em Tutoria Multissessão

Dissertação defendida no Programa de Pós–Graduação da Escola de Engenharia Elétrica e de Computação da Universidade Federal de Goiás como requisito parcial para obtenção do título de Mestre em Programa de Mestrado em Engenharia Elétrica e de Computação, aprovada em 22 de Agosto de 2009, pela Banca Examinadora constituída pelos professores:

Prof. Weber Martins, Ph.D.
Escola de Engenharia Elétrica e de Computação – UFG
Presidente da Banca

Prof. Lauro Eugênio Guimarães Nalini, Dr.
Departamento de Psicologia – PUC-Goiás

Prof. Alberto Nogueira de Castro Junior, PhD
Departamento de Ciência da Computação - UFAM

Prof. Leonardo da Cunha Brito, Dr.
Escola de Engenharia Elétrica e de Computação - UFG

Viviane Margarida Gomes

Bacharel em Engenharia de Computação pela Universidade Federal de Goiás (UFG). Colou grau e iniciou o mestrado em março de 2007. Durante sua graduação, foi colaboradora do Grupo de Pesquisa em Sistemas Inteligentes (PIRENEUS) (2003-2007), monitora da disciplina de Circuitos Elétricos (2004) e pesquisadora na iniciação científica - Projeto Fábrica Virtual, da Rede Internacional Virtual de Educação (RIVED) (2004). Durante o mestrado (2007-2009), na UFG/Grupo PIRENEUS, foi bolsista da CAPES e desenvolveu trabalho empírico sobre o controle inteligente de tempo livre em tutoria multissessão.

A quem dedicou a vida por mim, mamãe e papai.

Agradecimentos

Engenheira, pela formação acadêmica. Gosto exato!

Amiga-Filha-Neta-Sobrinha-Afilhada-Irmã-Prima-Madrinha-Tia, pelos amores que cultivei.

Papai e Mamãe, Donizete e Lourdes, quanto eu aprendi com vocês! Primeira grande lição: há um Papai e uma Mamãe do Céu que te ama muito. Amor especialmente revelado no mestrado. Meus queridos pais (sempre presentes), agradeço pela companhia que me acalentou e impulsionou a seguir em frente.

Grata, Weber, por tudo. "Vivi, p'ra saber como realmente eu sou, eu me imagino vivendo na China (numa cultura totalmente diferente da nossa)... Mesmo lá, não conseguiria ser desonesto, mau, falso." Que lindo! Longas conversas, reflexões inesquecíveis.

Querido Lauro, a "Querida Vivi" agradece por todo tempo dedicado a me orientar, sempre com alguma fala inusitada/engraçada.

Primeira aula do mestrado, antes mesmo da colação de grau. Colação: Diretor da EEEC[1], Reinaldo, anuncia que uma aluna já estava fazendo mestrado. Um amigo pergunta: "Quem é a doida?" Respondo: "Sou eu." Aos amigos da faculdade, meu "Muito obrigada!" pelo apoio e carinho (risos).

Um agradecimento especial a minha querida amiga Lá (Larissa). Esteve comigo o tempo todo, me encorajando a continuar, cuidando de mim quando precisei... Alguém com o coração bom, amigo, admirável!

Agradeço a todos da minha família: Vovó Maria José e vovô Atílio (por suportarem meu mau humor e me acompanharem dia após dia), meus irmãos Flávio (e Ádila, esposa) e Tarcízio (e Lívia, esposa, e Verônica, filhinha linda), meus queridos tios Zezinho, Geraldo, Ordália (e familiares), meus primos Karen, Betinha e Frank, Poliana e Rodrigo. Aos familiares por parte do papai, meus queridos tios/padrinhos Quecim (Deoclísio) e Doneci, tio Puíte (Messias) e primos Fernando (e Deise) e Patrícia (e Gabriel e Fernanda).

[1]EEEC é a Escola de Engenharia Elétrica e de Computação da Universidade Federal de Goiás - UFG.

No Grupo Pireneus, cada amigo me ajudou de alguma forma: Fê (Fernando), jovem (Victor), Fabrícia, Roberta, Kleber, Kono, Neto, Weder, Mcgill, Chico e a Carol. Eeehh! Carol, quanto você contribuiu para realização do experimento: com sua voz, com seu esforço na edição dos vídeos, na seleção da amostra e coleta de dados. Nunca me esquecerei do seu *post it* no meu PC logo após o preteste do sistema tutor. Apesar das muitas falhas do sistema, o seu recado era animador, de quem consegue ver o lado positivo de tudo.

Aos amigos João Felipe e Bruninho, agradeço pela constância em servir. Isso mesmo! Estiveram comigo na construção (em Java) do sistema tutor inteligente até o fim. Sem vocês, sinceramente, não sei se teria chegado onde cheguei (chegamos!)...

Na EEEC, aos amigos Wosley e Alexandre, suporte para avançar sempre. Aos servidores eficientes e super atenciosos, João, Dulce e Bosco. Ao pessoal da vigilância e da limpeza, Cícero, Divino, Valdenir, Sr. Antônio, Elias e Tininha, pela atenção dedicada.

No IFGoiás, amizades tão recentes e tão leais. Thays, quanto você me ajudou na coleta de dados. Valeu!!! Agradeço especialmente ao meu irmão-chefe Flávio e aos amigos Saulo (pela ajuda em Linux), Reinaldo, Cleiton, Antônio, Rômulo, Danielly, Cida, Kátia, Sara, Daniel e tantos outros.

Coleta de dados, eita etapa difícil! Vocês deram vida ao experimento: Casa da Juventude Pe. Burnier (Eduardo, Josiane e alunos), IFGoiás (Elymar, Joana, Márcia, Sr. Antônio e alunos), Guarda Mirim Inhumense (Bárbara Pessoni e alunos), Colégio Estadual Manoel Vilaverde (Cida Brito e alunos), Colégio Estadual Ary Valadão (Rita e alunos), Colégio Estadual Santa Rosa (Joilda e alunos).

Aos avaliadores da minha dissertação, professores Alberto Nogueira de Castro Júnior e Leonardo Brito. Na defesa, momento especialíssimo, a participação do Prof. Alberto (à distância, direto da Escócia) foi descontraída, desde o seu jeito de falar às brincadeiras do público quando a conexão caía. O Prof. Leonardo Brito me acompanhou em outros momentos, como professor, coordenador do mestrado e avaliador na qualificação, sempre prestativo.

Quantos agradecimentos! Ainda sim, provavelmente esqueci tantas outras pessoas que contribuíram para realização dessa conquista. Se você é uma delas, saiba que sempre que nos encontrarmos, recordarei a minha gratidão por você.

O mestrado chega ao fim. Mas a caminhada continua... Meus queridos companheiros, AMO vocês!

"Quem teve a idéia de cortar o tempo em fatias,
a que se deu o nome de ano,
foi um indivíduo genial.

Industrializou a esperança,
fazendo-a funcionar no limite da exaustão.

Doze meses dão para qualquer ser humano se cansar
e entregar os pontos.

Aí entra o milagre da renovação
e tudo começa outra vez, com outro número
e outra vontade de acreditar
que daqui para diante tudo vai ser diferente."

Carlos Drummond de Andrade,
Cortar o Tempo.

Resumo

Gomes, Viviane Margarida. **Controle Inteligente de Tempo Livre em Tutoria Multissessão**. Goiânia, 2009. **??**p. Dissertação de Mestrado. Escola de Engenharia Elétrica e de Computação, Universidade Federal de Goiás.

Sistemas Tutores Inteligentes são programas para prover instrução personalizada a partir de técnicas de Inteligência Computacional. Esta pesquisa propõe o controle inteligente de tempo livre (pausas) em tutoria multissessão. A estratégia de ensino apresenta a tutoria em módulos, com as seguintes etapas: 1) vídeo-aula, 2) exercício, 3) sugestão prática, 4) tempo livre e 5) exercício de revisão. Como parte do ambiente de aprendizagem, o tempo livre (etapa 4) pode contribuir para aumentar a retenção de conhecimento. Baseado no desempenho do aluno nos exercícios, o sistema proposto utiliza Aprendizagem por Reforço para controlar a duração do tempo livre. O agente inteligente toma decisões de acordo com a política definida pelo método Softmax. Entre os pontos relevantes do algoritmo, destacam-se o valor inicial otimizado das ações, a implementação incremental e o ajuste da temperatura (parâmetro da distribuição de Gibbs) para a seleção de ação. Dois grupos de estudantes participaram da coleta de dados. O grupo experimental (com controle inteligente do tempo livre) foi comparado ao grupo controle (onde a decisão pertence ao próprio estudante). Nos grupos, o agente inteligente ou o aluno determina a ação a ser seguida, mais detalhadamente, diminuir, manter ou aumentar a duração do tempo livre. Por meio de estudo comparativo, a análise estatística dos dados mostrou ganhos significativos e equivalentes na retenção de conhecimento. Contudo, alunos do grupo experimental perceberam melhor o tempo livre como componente da estratégia de ensino.

Palavras–chave

Sistemas Tutores Inteligentes, Aprendizagem por Reforço, Tempo Livre

Abstract

Gomes, Viviane Margarida. **Intelligent Control of Free Time in Multi-session Tutoring**. Goiânia, 2009. **??**p. MSc. Dissertation. Escola de Engenharia Elétrica e de Computação, Universidade Federal de Goiás.

Intelligent Tutoring Systems are softwares to provide customized instruction by using techniques of Computational Intelligence. This research proposes the intelligent control of free time (break interval) in multi-session tutoring. The teaching strategy employs tutoring modules with the following steps: 1) video class, 2) exercise, 3) practical suggestion, 4) free time, and 5) revision exercise. As part of the learning environment, free time (step 4) can contribute to increase the knowledge retention. Based on the student performance in exercises, the proposed system uses Reinforcement Learning to control free time durations. The intelligent agent decides according to the policy that has been indicated by the Softmax method. Among the relevant points of this algorithm, it can be highlighted the optimistic initial values, the incremental implementation and the temperature adjustment (Gibbs distribution parameter) to the selection of action. Two student groups have participated of data collection. The experimental group (with intelligent control) has been compared to the control group (where decisions belong to the student). In the groups, the intelligent agent or the student determines the action that will be followed or, in more detail, if free time will be shorter, longer or maintained. In comparison, statistical data analysis have shown significant and equivalent gains in knowledge retention. However, students from experimental group have realized more accurately the role of free time as a component of the teaching strategy.

Keywords

Intelligent Tutoring Systems, Reinforcement Learning, Free Time

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

1.1 Tema

Esta pesquisa investiga o tema "Sistemas Tutores Inteligentes" (STI), programas computacionais para prover instrução personalizada [Leung e Li 2007]. A partir da interação homem-máquina, técnicas de Inteligência Computacional utilizam-se de características do aprendiz para adaptar dinamicamente o conteúdo e/ou o estilo de ensino [Murray 1999].

Sistemas Tutores Inteligentes, conhecidos ainda como Instrução Inteligente Assistida por Computador (do Inglês, *Intelligent Computer Assisted Instruction, ICAI*), agregam os benefícios do ensino individualizado, onde o aluno progride no seu próprio ritmo [Skinner 1972], aos recursos tecnológicos para tratar o aprendiz de forma única [Martins et al. 2004].

Tradicionalmente, os STIs (ou ITSs, do Inglês, *Intelligent Tutoring Systems*) possuem quatro componentes principais: a base de conhecimento do domínio, ou módulo especialista, o modelo do estudante, ou módulo de diagnóstico do estudante, o módulo pedagógico, ou tutor, e o módulo de comunicação [Sarrafzadeh et al. 2008] [Burns e Capps 1988] (ver Figura 1.1).

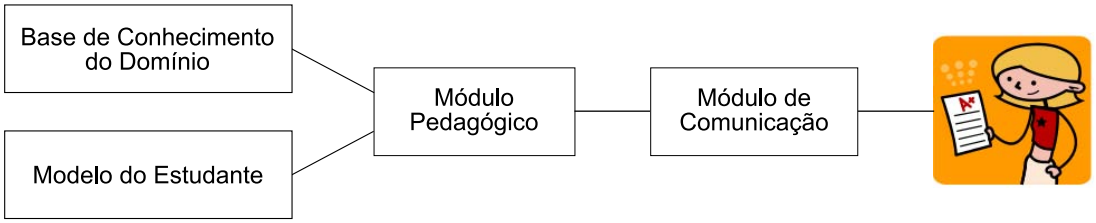

Figura 1.1: *Principais componentes de um sistema tutor inteligente (Fonte: [Sarrafzadeh et al. 2008]).*

Segundo [Sarrafzadeh et al. 2008], o módulo pedagógico escolhe estratégias de ensino apropriadas para o STI aplicar, baseado no modelo do estudante, o qual armazena informação dos aprendizes individualmente. Essas estratégias, aplicadas ao conhecimento do domínio, geram um subconjunto

de conhecimento a ser apresentado para o aprendiz usando o modelo de comunicação, ou seja, a interface entre o aprendiz e o STI. Assim que o aluno responde ao sistema, o modelo do estudante é atualizado e o ciclo repete.

Neste trabalho, o sistema proposto emprega tempo livre (pausas) entre sessões de tutoria. Como ponto crucial de investigação, acredita-se que uma atuação de forma inteligente no controle do tempo livre pode levar a bons resultados. No STI disposto nesta pesquisa, o modelo do estudante contém medidas do desempenho do aluno, a base de conhecimento de domínio apresenta o material instrucional e o método de avaliação do aprendiz, o módulo pedagógico realiza o controle (mantém ou modifica a duração) do tempo livre e, por fim, o módulo de comunicação interage com o estudante por meio da interface gráfica do *software*.

1.2 Justificativa

O processo de ensino-aprendizagem necessita adaptar-se ao mundo contemporâneo. Longas aulas expositivas mostram-se ineficientes, exigindo novos paradigmas de tutoria. Cada vez mais, adolescentes permanecem diante do computador em atividades de lazer (jogos, vídeos) e relacionamento (chats, orkuts) [Subrahmanyam et al. 2001]. Além de prover entretenimento, os recursos computacionais são poderosas ferramentas para o ensino.

Segundo [Eberspächer e Kaestner 2009], com o uso de Sistemas Tutores Inteligentes (*softwares* educacionais), os estudantes avançam para assuntos mais complexos em um terço do tempo gasto com a metodologia convencional e apresentam 40% de aumento no desempenho comparado ao ensino em sala de aula.

Embora existam meios eficientes para instrução, fala-se atualmente em Era da Ignorância (em oposição à Era do Conhecimento), por haver relativamente pouca geração de conhecimento comparada ao grande volume de informações. Este cenário sugere o uso das Tecnologias da Informação e Comunicação (TICs) na otimização de processos, além de tornarem-se tema de estudos (capacitação e formação).

De acordo com o Centro de Estudos sobre as Tecnologias da Informação e da Comunicação - CETIC.br [Barbosa 2008], do Comitê Gestor da Internet no Brasil (CGI.br), 60% dos usuários[1] de computador carecem de ha-

[1]Percentual sobre o total de pessoas que já utilizaram computador e declararam possuir alguma habilidade.

bilidades tecnológicas suficientes para o mercado de trabalho.

O Índice Brasil para Convergência Digital (IBDC) apresenta um quadro ruim quanto à inclusão digital, referente ao acesso às TICs e ao desenvolvimento de competências para utilizá-las. Nelson Wortsman, diretor de convergência digital da BRASSCOM[2], afirma que "a educação é o lado mais frágil do índice" [Folha 2008].

A exclusão digital mantém as pessoas alheias ao mundo revelado pela Internet. Conforme indicadores do CETIC.br, a principal causa apontada pelos respondentes para nunca terem acessado à Internet é a falta de habilidade com o computador/Internet, como mostra a Figura 1.2. Este gráfico apresenta o percentual sobre o total de pessoas que nunca acessaram à rede mundial de computadores, mas usaram o computador.

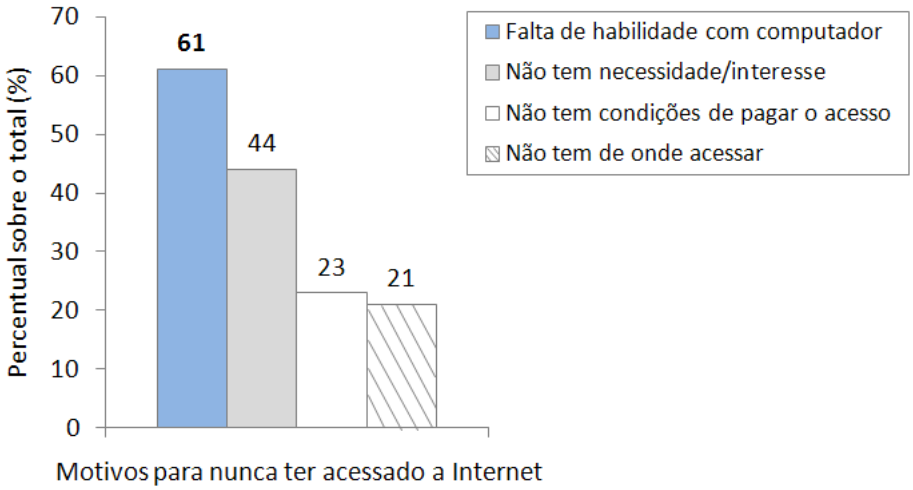

Figura 1.2: *Motivos pelos quais nunca utilizou a Internet (Fonte: CETIC.br).*

Entre algumas iniciativas para inclusão digital, o Governo Federal incentiva o uso de softwares livres, em programas como: Casa Brasil, Computador para Todos, Maré - Telecentros da Pesca, Pontos de Cultura - Cultura Digital e Quiosque do Cidadão [Portal 2009]. Segundo a Free Software Foundation [Foundation 2009], software livre é o sistema de computador que fornece ao usuário a liberdade para usá-lo, distribuí-lo, estudar seu código-fonte e modificá-lo.

O programa "Computador para Todos" favorece a aquisição de computadores com, obrigatoriamente, *softwares* livres. Apesar do apoio governa-

[2]Associação Brasileira de Empresas de Tecnologia da Informação e Comunicação (Brazilian Association of Information Technology and Communication Companies). A BRASSCOM é responsável pelo IBDC (Índice Brasil para Convergência Digital).

mental, a ausência desse equipamento nos domicílios deve-se, em parte, pelo pouco conhecimento sobre Informática. Com 29% sobre o total de residências sem acesso ao computador, a falta de habilidade ocupa o terceiro lugar no quadro das causas, como mostra a Figura 1.3.

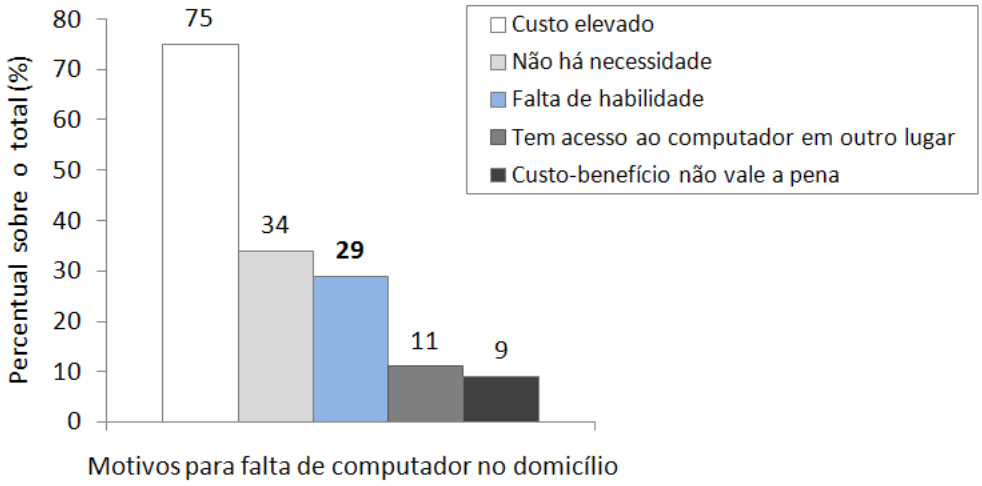

Figura 1.3: *Motivos para a falta de computador no domicílio (Fonte: CETIC.br).*

Baseado na realidade atual, esta pesquisa pretende contribuir para capacitação de adolescentes no uso do computador. Os alunos poderão aprender conceitos sobre Microinformática, como Linux (sistema operacional livre), no próprio Linux. Aprender sobre tecnologia a partir de recursos tecnológicos, como Sistemas Tutores Inteligentes. A ideia básica é prover pausas para prática ou entretenimento após cada etapa de estudo, com controle inteligente desses tempos livres[3].

A pausa pode favorecer a aprendizagem sob distintos pontos de vista: construtivista e comportamentalista. Para construtivistas, o tempo livre possibilita postura ativa do aprendiz na construção do conhecimento, pois fica exposto ao ambiente e explora as suas funcionalidades. No ponto de vista comportamental, por sua vez, a pausa reforça o tempo dedicado ao estudo (sessão de tutoria), sendo a condição de liberdade reforçadora para o aluno.

O sistema proposto atinge outras faixas etárias e novos conteúdos, além de ser provido facilmente em vários lugares do Brasil e, até mesmo, do mundo. De forma geral, a implementação de *softwares* educacionais é um investimento altamente proveitoso, afinal, uma vez prontos, tais programas

[3]"Tempo livre" é o termo usado para as pausas entre as sessões de tutoria. Durante a pausa, o aluno está livre para usar o computador conforme sua vontade.

possuem alcance considerável, replicação em alta escala em mídia (CDs, DVDs, pendrives) ou pela Internet. No caso de Sistemas Tutores Inteligentes, há vantagens no seu uso em relação à instrução tradicional, evidência de que a Educação deve acompanhar as mudanças da sociedade.

1.3 Problema e Hipóteses

Historicamente, os Sistemas Tutores Inteligentes possuem etapas consecutivas para o ensino do conteúdo. Como um novo recurso, esta pesquisa propõe a inserção de tempo livre entre as fases da tutoria (ver a Figura 1.4). Baseado no cenário de tutoria multissessão, este trabalho apresenta estudo confirmatório sobre a solução provisória para o problema: "Como controlar tempo livre em Sistemas Tutores Inteligentes?".

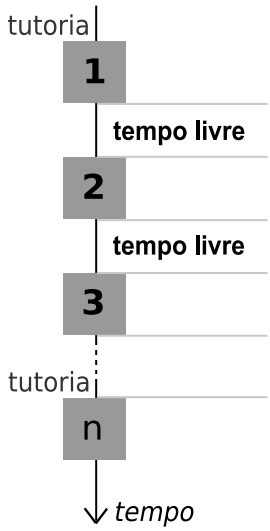

Figura 1.4: *Cenário da tutoria.*

Com base nas Teorias da Aprendizagem por Reforço, em Inteligência Computacional, e da Análise Experimental do Comportamento, a partir do princípio de Premack, acredita-se que o uso de **Aprendizagem por Reforço** para **controlar o tempo livre** aumenta a retenção de conhecimento (hipótese básica). O sistema proposto deve aprender a ação adequada ao aluno no controle das pausas, situação colocada à prova por meio de método experimental [Cohen 1995]. Caso a hipótese básica seja fortalecida, as hipóteses secundárias levantadas afirmam menor tempo total gasto na atividade e maior satisfação do aluno.

1.4 Visão Geral dos Capítulos

O Capítulo 2 apresenta a **Fundamentação Teórica**: Aprendizagem por Reforço e Análise Experimental do Comportamento. Entre muitas técnicas, Aprendizagem por Reforço destaca-se como micromundo da Inteligência Computacional, sendo utilizada no sistema proposto (Capítulo 3). Considerada como uma Ciência do Comportamento, a Análise Experimental do Comportamento fornece conteúdo teórico para fundamentar o uso de tempo livre em tutoria multissessão.

O Capítulo 3 dedica-se à descrição do **Sistema Proposto**. O controle inteligente utiliza Aprendizagem por Reforço para adaptar a duração da pausa ao estudante. O cenário de tutoria alterna tempo no sistema tutor inteligente e tempo livre, para descanso ou para atividades práticas relacionadas ao conteúdo. Para convergir, o agente inteligente toma decisões baseado no método Softmax, detalhado no Capítulo 2. Como pontos importantes do algoritmo, destacam-se a definição do valor inicial das ações, a implementação incremental e o ajuste da temperatura, parâmetro do método de seleção de ação.

O Capítulo 4 apresenta o **Experimento e Resultados** da validação empírica. O experimento constitui a implementação do sistema proposto, instanciando elementos como as etapas da tutoria, os estados, ações e recompensas, a temperatura do método Softmax e outros. As condições de coleta de dados, ambiente tecnológico e amostra contextualizam o cenário da pesquisa. A partir dos dados registrados, a análise e a interpretação (desses dados) fornecem os resultados, com evidências sobre a capacidade de "ensinar" dos sistemas tutores, ou seja, observação de diferenças significativas entre as notas iniciais e finais dos alunos.

No Capítulo 5, a **Conclusão**, as seções "Considerações gerais", "Principais contribuições" e "Trabalhos futuros" abordam questões sobre a pesquisa como um todo, como a sua relevância científica e social, com sugestões para novas investigações. Por se tratar de um trabalho pioneiro, as observações servem de base para que pesquisadores posssam dar sequência ao estudo do tempo livre em estratégias de ensino.

CAPÍTULO 2

Fundamentação Teórica

O presente capítulo apresenta a fundamentação teórica deste trabalho: Aprendizagem por Reforço e Análise Experimental do Comportamento. Juntas, estas teorias formam a base do sistema proposto (Capítulo 3), o controle inteligente de tempo livre em tutoria multissessão.

2.1 Aprendizagem por Reforço

Em diversas áreas, o processo de aprendizagem é estudado. Em Ciência da Computação, Inteligência Computacional pesquisa métodos e técnicas para simular a capacidade humana de resolver problemas. Nesse sentido, paradigmas de aprendizagem são bases no desenvolvimento de sistemas inteligentes e possuem estudo detalhado no subcampo de Inteligência Computacional conhecido por Aprendizagem de Máquina.

Segundo [Russell e Norvig 2003], aprendizagem é a melhoria das habilidades do agente para agir no futuro. No contexto de Aprendizagem de Máquina, o agente possui dois elementos importantes: o de desempenho e o de aprendizagem. O elemento de desempenho decide quais ações serão executadas, enquanto o elemento de aprendizagem modifica o elemento de desempenho para que ele possa tomar decisões melhores.

O cenário de interação entre os elementos de desempenho e de aprendizagem acontece através de realimentações (do Inglês, *feedbacks*) apropriadas. A natureza do problema para o agente depende da definição de realimentação disponível. Usualmente três tipos de aprendizagem são usados: Supervisionada, Não-supervisionada e Aprendizagem por Reforço [Russell e Norvig 2003].

Haykin [Haykin 2001] apresenta os tipos de aprendizagem em duas classes: com professor e sem professor. Tal classificação define a diferença principal entre os tipos, ou seja, possuir ou não conhecimento prévio do ambi-

ente. De forma sintética, os paradigmas de aprendizagem serão apresentados abaixo.

A forma de aprendizagem com professor é:

- **Supervisionada**: o conhecimento do professor é representado por exemplos de entrada e saída [Haykin 2001]. A partir destes exemplos, o algoritmo aprende a associar cada vetor de entrada x ao seu vetor de saída y correspondente, num processo conhecido como treinamento [Kasabov 1998]. A aprendizagem acontece quando o algoritmo responde de forma satisfatória a novas situações, processo conhecido como generalização.

Sem professor, existem os tipos de aprendizagem:

- **Não-supervisionada**: o algoritmo aprende padrões do conjunto de entradas, visto que não são fornecidos valores de saída específicos [Russell e Norvig 2003].
- **Aprendizagem por Reforço**: a interação contínua com o ambiente proporciona o mapeamento de situações para ações [Sutton e Barto 1998]. A ausência de professor é compensada pela presença de um crítico, que converte um sinal de reforço primário recebido do ambiente em um sinal de reforço de melhor qualidade, denominado sinal de reforço heurístico, como mostra a figura abaixo [Haykin 2001].

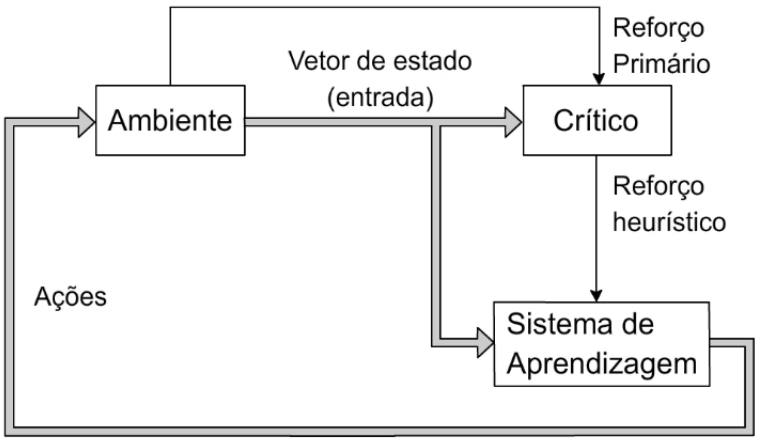

Figura 2.1: *Diagrama em blocos da Aprendizagem por Reforço (Fonte: [Haykin 2001]).*

Este trabalho utiliza a técnica Aprendizagem por Reforço para prover ensino personalizado. Logo tal paradigma será estudado de forma detalhada, com enfoque na história, nos conceitos básicos, como funções-valor e política, e

nos métodos de solução (implementação), como a Aprendizagem por Diferença Temporal.

2.1.1 Breve Histórico

Os primeiros trabalhos de Aprendizagem por Reforço foram realizados no início da Inteligência Computacional, Cibernética, e contou com conhecimento de outras áreas, como Estatística, Psicologia, Neurociência e Ciência da Computação [Kaelbling, Littman e Moore 1996]. Investigações sobre aprendizagem por tentativa e erro foram publicadas, em 1954, por Minsky e por Farley e Clark. Em 1960, pela primeira vez, os termos "reforço" e "aprendizagem por reforço" foram altamente usados na literatura de Engenharia [Sutton e Barto 1998].

A história de **Aprendizagem por Reforço** é vista sob a perspectiva de três linhas distintas, segundo [Sutton e Barto 1998]. Duas delas foram desenvolvidas independentemente, aprendizagem por **tentativa e erro** e problema de **controle ótimo**. As exceções destas abordagens levam a outra: métodos de **diferença temporal**.

Aprendizagem por tentativa e erro foi inicialmente verificada no campo da Psicologia, nas primeiras décadas do século XX. Estudos sobre aprendizagem animal mostraram o fortalecimento de ações com consequências boas. E, de forma contrária, ações de efeito ruim eram enfraquecidas, ou seja, teriam menor chance de acontecer em situações similares. Nesse sentido, dois aspectos importantes da aprendizagem por tentativa e erro são evidenciados: seleção e associação. Após várias tentativas, é possível *selecionar* as melhores ações, tendo em vista as consequências delas, e *associá-las* a situações particulares em que aconteceram. No caso de Aprendizagem por Reforço, estes aspectos da aprendizagem por tentativa e erro são essenciais.

Na área de Inteligência Computacional, Aprendizagem por Reforço destaca-se por implementar tanto *feedback* avaliativo quanto instrutivo. Existem técnicas puramente avaliativas, como Algoritmos Genéticos, ou puramente instrutivas, como Redes Neurais Artificiais. Aprendizagem por Reforço agrega os benefícios da avaliação à instrução. Ao avaliar cada ação aplicada, o agente guarda informações importantes para decisões futuras. Sutton e Barto [Sutton e Barto 1998] caracterizam esse processo como busca e memória.

Durante as décadas de 1960 e 1970, a pesquisa sobre aprendizagem por tentativa e erro tornou-se rara. Em partes, devido a confusões, existentes ainda hoje, relacionadas à idéia de tentativa e erro para a aprendizagem

supersionada, por basear-se em erro para ajuste dos pesos de uma rede neural, por exemplo. No entanto, essa confusão é esclarecida quando evidencia-se o caráter essencial da aprendizagem por tentativa e erro conhecido por seleção, tipicamente usado em Aprendizagem por Reforço.

Em relação à teoria de controle, a programação dinâmica destaca-se como uma classe de métodos eficientes para resolver problemas estocásticos de controle ótimo. Desde a década de 1950, ela tem sido altamente aplicada em comparação a outros métodos gerais. Para muitos, programação dinâmica não faz parte de técnicas de Aprendizagem por Reforço, pois necessita de conhecimento completo do sistema a ser controlado. Porém, para [Sutton e Barto 1998], muitos métodos de programação dinâmica são incrementais e iterativos e, assim como métodos de aprendizagem, eles gradualmente alcançam a resposta correta através de sucessivas aproximações.

Com origem na psicologia também, mais precisamente, na noção de reforçadores secundários, destacam-se os métodos de diferença temporal. Um reforçador secundário é um estímulo associado a um reforçador primário, aquele que naturalmente produz efeito positivo, como água, alimento, afeto. Estes métodos de aprendizagem são distintos por serem dirigidos pela diferença temporal entre estimativas sucessivas da mesma quantidade.

A partir da década de 1980, as três linhas trabalham juntas para produzir o moderno campo da Aprendizagem por Reforço. Aplicações em robótica [Matignon e Fort-Piat 2007], educação [Martins et al. 2007], saúde [Araújo 2000] [Kalidindi e Bowman 2007], jogos [Frénay e Saerens 2009] evidenciam a flexibilidade da Aprendizagem por Reforço.

2.1.2 Conceitos básicos

Aprendizagem por Reforço é o conjunto de técnicas inteligentes baseadas na interação entre agente e ambiente. O contato permite ao **agente** (algoritmo) aprender as ações adequadas ao **ambiente** (pessoa, espaço físico). Segundo [Kaelbling, Littman e Moore 1996], Aprendizagem por Reforço é um paradigma adequado ao problema enfrentado por agentes que devem aprender a comportar-se em ambientes dinâmicos através de interações de tentativa e erro.

A estrutura (do Inglês, *framework*) da Aprendizagem por Reforço fundamenta-se na relação entre estado-ação-recompensa. **Estado** é o contexto de interação entre ambiente e agente a cada momento. O agente interage com o ambiente por meio de alguma **ação**. Em resposta à ação aplicada, o

ambiente fornece uma **recompensa**, servindo de base para avaliar a escolha do agente. Ao longo da interação, o agente aprende escolher boas ações (ou pares estado-ação) para atingir seu **objetivo**: maximizar o valor total das recompensas recebidas.

Agente e ambiente interagem em uma sequência de passos discretos. A cada passo, o agente escolhe alguma ação a partir da **política**, denotada por π. Política é a regra estocástica pela qual o agente seleciona ações como uma função de estados [Sutton e Barto 1998]. Em outras palavras, política é a distribuição de probabilidades das ações em cada estado. Matematicamente, $\pi_t(s, a)$ é a probabilidade de $a_t = a$ se $s_t = s$.

A probabilidade $\pi_t(s, a)$ resulta de uma **função-valor** combinada ao critério específico da política π_t. A cada estado, funções-valor designam o **retorno** esperado daquele estado, ou par estado-ação. O retorno R_t é a função de recompensas futuras que o agente busca maximizar, como mostra a equação abaixo, onde T é o passo final no tempo.

$$R_t = r_{t+1} + r_{t+2} + r_{t+3} + ... + r_T \tag{2-1}$$

Uma especificação da interface dos passos (entre agente e ambiente) define uma tarefa. Existem **tarefas episódicas** e **contínuas**. Episódios possuem passo final definido. Ao contrário, tarefas contínuas são caracterizadas pela interação entre agente e ambiente indefinidamente, sem término estabelecido. Nesse caso, o retorno deve ter recompensas descontadas, ou seja, valores presentes de recompensas futuras. A taxa de desconto γ, sendo que $0 \leq \gamma \leq 1$, minimiza a influência de recompensas incertas, distantes do momento atual, conforme a Equação 2-2.

$$R_t = r_{t+1} + \gamma.r_{t+2} + \gamma^2.r_{t+3} + ... = \sum_{k=0}^{\infty} \gamma^k.r_{t+k+1} \tag{2-2}$$

Se o sinal de estado do ambiente sintetiza de forma compacta o passado sem degradar a habilidade de predizer o futuro, diz-se que o ambiente satisfaz a **Propriedade de Markov**. Com esta propriedade assegurada, o ambiente define um **Processo de Decisão de Markov** (MDP, do Inglês, Markop Decision Process). Um MDP finito possui conjuntos finitos de estado e ação, sendo a restrição da maioria das teorias atuais de Aprendizagem por Reforço.

Para tomar boas decisões, o agente busca uma **política ótima** π^*, resultado de funções-valor ótimas (de estado V^* e de ação Q^*), ou seja, com maior retorno esperado. Para facilitar a determinação da política ótima, as

funções-valor ótimas devem satisfazer condições de consistência, conhecidas como *Equações de Otimilidade de Bellman*.

O uso de funções-valor distingue métodos de Aprendizagem por Reforço de métodos evolucionários, que avaliam políticas inteiras em sucessivos passos (gerações) para encontrar a política ótima. Portanto as funções-valor são essenciais para busca eficiente no espaço de políticas, tornando-se características centrais em Aprendizagem por Reforço, como detalha a próxima seção.

2.1.3 Funções-valor

As funções-valor são funções para estimar valores de estados ou ações para política π. Essas funções estimam quão bom é desempenhar certa ação num estado, ou seja, o retorno esperado para a ação. O valor (estimativa) $Q_t(a)$ de uma ação a pode ser a média das recompensas recebidas até o momento atual t. Se uma ação a foi aplicada k_a vezes até o instante t e recebeu as recompensas r_1, r_2,..., r_{k_a}, então o seu valor pode ser estimado pela equação abaixo.

$$Q_t(a) = \frac{r_1 + r_2 + ... + r_{k_a}}{k_a} \tag{2-3}$$

Para $k_a = 0$, $Q_t(a)$ assume o valor inicial $Q_0(a)$ estabelecido. Se $k_a \to \infty$, então $Q_t(a)$ converge para o valor verdadeiro da ação. Nesse caso, a ação de maior valor a^* possui função-valor ótima $Q^*(a)$, onde $Q^*(a) = Q_t(a^*) = max_a Q_t(a)$.

Para MDPs, a **função estado-valor** V^π (como mostra a Equação 2-4) fornece o valor de um estado s, denotado por $V^\pi(s)$, seguindo a política π. Para determinar o valor $Q^\pi(s,a)$ de uma ação a, a **função ação-valor** Q^π calcula o retorno esperado começando de s, aplicada a ação a, e seguindo a política π, conforme Equação 2-5. O valor esperado é denotado por E_π.

$$V^\pi(s) = E_\pi\{R_t | s_t = s\} = E_\pi\{\sum_{k=0}^{\infty} \gamma^k . r_{t+k+1} | s_t = s\} \tag{2-4}$$

$$Q^\pi(s,a) = E_\pi\{R_t | s_t = s, a_t = a\} = E_\pi\{\sum_{k=0}^{\infty} \gamma^k . r_{t+k+1} | s_t = s, a_t = a\} \tag{2-5}$$

Graficamente, [Sutton e Barto 1998] mostram a atualização do valor do estado (ou estado-ação) em diagramas de *backup* (palavra inglesa), ou seja,

que retratam a experiência do agente ao representar o relacionamento entre estados e ações (Figura 2.2).

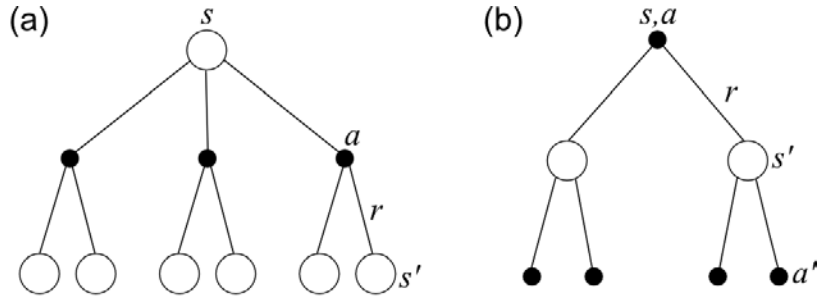

Figura 2.2: *Diagramas de backup para (a) V^{π} e (b) Q^{π} (Fonte: [Sutton e Barto 1998]).*

As funções-valor são inicializadas com um valor inicial, $V_0(s)$ e $Q_0(s, a)$, definido arbitrariamente ou de forma otimizada. A cada passo, o agente atualiza o valor do estado (ou estado-ação), de forma incremental, por exemplo. Estes dois pontos, valor inicial otimizado e implementação incremental, são detalhados abaixo.

Valor Inicial Otimizado

O valor inicial de cada estado (ou estado-ação) pode contribuir para aprendizagem computacional acurada. Entende-se por valor inicial otimizado a estimativa que proporciona, a longo prazo, valores mais fiéis das ações. Em muitos casos, a estimativa inicial encoraja a exploração extensiva (detalhada na próxima seção), ou seja, a seleção de ações desconhecidas. Tal prática maximiza o total de recompensas recebidas, objetivo do paradigma da Aprendizagem por Reforço.

Em Estatística, as estimativas iniciais são chamadas *bias* (do Inglês, "tendência"). Para métodos com média amostral (conforme Equação 2-3), a *bias* desaparece à medida que as ações são selecionadas. Ao contrário, em métodos com a constante α, parâmetro *step-size*[1], a *bias* permanece na função-valor, diminuindo sua influência com o tempo (passos).

Implementação Incremental

A implementação incremental atualiza o valor dos estados ou ações como função do valor anterior. Em métodos de ação-valor, as estimativas podem resultar de médias amostrais das recompensas observadas (ver a Equa-

[1]Parâmetro TamanhoPasso (do Inglês, *step-size*). Em problemas não-estacionários, ou seja, onde o valor verdadeiro de cada ação muda com o tempo, utiliza-se α constante.

ção 2-3). À medida que o agente registra as recompensas das ações, cresce a demanda por memória avaliativa e requisitos computacionais. Atualizar as estimativas de forma incremental evita esse problema, como mostra a seguinte regra:

$$EstimativaNova \leftarrow EstimativaVelha + TamanhoPasso[Alvo - EstimativaVelha] \tag{2-6}$$

O parâmetro TamanhoPasso é denotado pelo símbolo α por [Sutton e Barto 1998]. A expressão $[Alvo - EstimativaVelha]$ é um erro na estimativa, sendo que o $Alvo$ indica a direção para o agente seguir. Pela regra (2-6), o valor de uma ação $Q_t(a)$ obtido pela média amostral das recompensas recebidas pode ser atualizado conforme a equação abaixo.

$$Q_t = Q_{t-1} + \frac{1}{t}[r_t - Q_{t-1}] \tag{2-7}$$

Esta implementação requer memória apenas para Q_t e t e pequena computação para cada nova recompensa. Os métodos de Aprendizagem por Reforço mais famosos, como o ***Q-learning***, são totalmente incrementais, conforme explica a Seção 2.1.5, Métodos de Implementação.

2.1.4 Política

A política, como dito anteriormente, é a regra estocástica pela qual o agente seleciona ações como uma função de estados. Matematicamente, $\pi_t(s, a)$ é a probabilidade de $a_t = a$ se $s_t = s$. Métodos de seleção de ação definem $\pi_t(s, a)$ baseado no valor $Q_t(s, a)$. Os métodos mais comuns são:

- *Greedy,*
- ϵ-*Greedy,*
- Softmax.

O método *Greedy* ("guloso") seleciona a ação de maior valor, com $Q_t(a^*) = max_a Q_t(a)$. Neste caso, a política $\pi_t(s, a^*) = 1$, enquanto as probabilidades das outras ações são iguais a zero. Dessa forma, o agente maximiza apenas as recompensas imediatas, impedindo a busca por ações melhores. Como solução a este problema, os métodos ϵ-*Greedy* e Softmax exploram extensiva e intensivamente as ações.

A exploração extensiva (do Inglês, *exploration*) contribui para maximizar o total de recompensas a longo prazo, ao escolher ações com valores menores de estimativas ação-valor. O método ϵ-*Greedy* apresenta uma pequena

probabilidade ϵ de escolher alguma ação diferente de a^* (de maior valor estimado). Qualquer ação, exceto a^*, possui a mesma chance de ser escolhida, ϵ. Portanto para a ação a^*, a política $\pi_t(s, a^*) = 1 - \epsilon$.

Para otimizar as decisões do agente, o método Softmax diferencia as probabilidades das ações de acordo com suas estimativas de ação-valor, $Q_t(s, a)$. Quanto maior o valor da ação, maior a probabilidade de ser escolhida. Para cálculo das probabilidades, Softmax utiliza, por exemplo, a distribuição de Gibbs, ou Boltzmann, como mostra a equação abaixo.

$$\pi_t(a) = \frac{e^{\frac{Q_t(a)}{\tau}}}{\sum_{b=1}^{n} e^{\frac{Q_t(b)}{\tau}}} \tag{2-8}$$

O balanceamento entre exploração extensiva e intensiva (do Inglês, *exploitation*), a curto prazo, realiza-se mediante o ajuste do parâmetro τ, chamado de "temperatura". Inicialmente, τ apresenta valor alto, a fim de tornar as ações equiprováveis, e gradativamente diminui, até atingir a diferenciação desejada entre as ações. Quando $\tau \to 0$, o Softmax comporta-se como o método *Greedy*.

Existem, ainda, os métodos de estimação de intervalos, onde estimativas de incerteza do valor das ações encorajam a exploração extensiva. Se uma ação foi escolhida poucas vezes, seu valor verdadeiro pode ser melhor do que o da ação a^*. Nesse sentido, métodos de estimação de intervalos favorecem a seleção de ações pouco conhecidas. Na prática, esses métodos são problemáticos, devido à complexidade de testes estatísticos usados para estimar o intervalo de confiança.

2.1.5 Métodos de Implementação

Em Aprendizagem por Reforço, a interação entre agente e ambiente objetiva resolver um problema. As classes fundamentais de métodos de solução desse paradigma são:

- Programação Dinâmica,
- Métodos de Monte Carlo,
- Aprendizagem por Diferença Temporal.

Conforme a caracterização do problema, define-se o método para solucioná-lo. A modelagem da situação fornece informações ao agente sobre o nível de conhecimento, completo ou incompleto, do ambiente dinâmico. Métodos de Programação Dinâmica exigem completo e acurado modelo do ambiente. Ao contrário, métodos de Monte Carlo e de Aprendizagem por Diferença

Temporal aprendem diretamente da experiência, mesmo em ambiente totalmente desconhecido.

Os métodos de **Programação Dinâmica** (do Inglês, *Dynamic Programming, DP*) são matematicamente bem desenvolvidos, usados para resolver Processos de Decisão de Markov (MDPs, do Inglês, *Markov Decision Process*). Segundo [Russell e Norvig 2003], um MDP é definido por três componentes:

- Estado inicial: s_0
- Modelo de transição: $T(s, a, s')$
- Função-recompensa: $R(s)$

Sendo o modelo de transição $T(s, a, s')$ comparado a uma grande tabela tridimensional contendo probabilidades (de alcançar, dependendo apenas de s, o estado s' se a ação a for executada no estado s).

Métodos de Programação Dinâmica atualizam as estimativas dos valores de estados baseadas em estimativas de valores de estados sucessores, com destaque para a iteração de política e a iteração de valor, resultado da união de recursos computacionais de avaliação e melhoria da política. A idéia geral de processos que interagem a fim de avaliar e melhorar a política recebe o nome de Iteração de Política Generalizada (do Inglês, *Generalized Policy Iteration, GPI*), usada em praticamente todos os métodos de Aprendizagem por Reforço (ver Figura 2.3) [Sutton e Barto 1998].

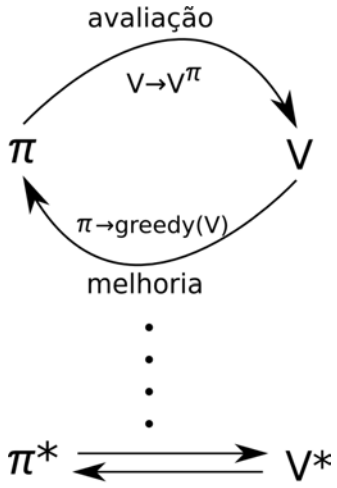

Figura 2.3: *Iteração de Política Generalizada: funções-valor e política interagem até se tornarem ótimas (Fonte: [Sutton e Barto 1998]).*

Diferentemente da Programação Dinâmica, métodos de **Monte Carlo** caracterizam-se pela aprendizagem direta a partir da interação com o ambiente, sem um modelo prévio. E atualizam as estimativas dos valores dos estados com base em episódios amostrais, ao invés de utilizar outras estimativas. Basicamente, métodos de Monte Carlo calculam o retorno a cada passo seguindo a primeira ocorrência da ação ou par estado-ação. Depois atualizam os valores das ações como a média dos retornos amostrais. A média torna-se uma boa aproximação para o valor, afinal o valor de estado (ou ação) é o retorno esperado.

Combinando as vantagens da Programação Dinâmica e de Monte Carlo, a **Aprendizagem por Diferença Temporal** (do Inglês, *temporal-difference, TD, learning*) destaca-se como uma idéia central e singular em Aprendizagem por Reforço. Os métodos de Aprendizagem por Diferença Temporal podem aprender diretamente da experiência sem um modelo do ambiente e atualizam as estimativas dos valores dos estados (ou ações) baseados em outras estimativas. Além disso, esses métodos necessitam apenas de uma quantia mínima de computação por serem totalmente incrementais.

Em Aprendizagem por Diferença Temporal, o balanceamento entre exploração extensiva e intensiva realiza-se em duas classes de métodos: política-ligada (do Inglês, *on-policy*) e política-desligada (do Inglês, *off-policy*). Métodos *on-policy*, como o Sarsa, dependem da política para aproximar o valor Q, função ação-valor aprendida, para Q^* (função ação-valor ótima). Ao contrário, para convergir corrretamente, os métodos de Diferença Temporal *off-policy* requerem apenas que todos os pares estado-ação sejam atualizados, com destaque para o ***Q-Learning*** (ver Figura 2.4).

Inicialize $Q(s,a)$ arbitrariamente
Repita (para cada episódio):
  Inicialize s
  Repita (para cada passo do episódio):
    Escolha a de s usando política derivada de Q (por ex., ε-greedy)
    Tomada ação a, observe r, s'
    $Q(s,a) \leftarrow Q(s,a) + \alpha [r + \gamma . max_{a'} Q(s',a') - Q(s,a)]$
    $s \leftarrow s'$;
  Até que s seja terminal

Figura 2.4: *Q-Learning: Um algoritmo de controle TD off-policy (Fonte: [Sutton e Barto 1998]).*

O algoritmo *Q-Learning*, desenvolvido por Watkins, constitui uma das mais importantes descobertas da Aprendizagem por Reforço, sendo enormemente utilizado na atualidade [Sutton e Barto 1998]. Segundo [Watkins e Dayan 1992], *Q-learning* pode ser visto como um método de Programação Dinâmica assíncrono, pois o agente tem a capacidade de aprender a agir otimamente em domínios Markovianos pela experimentação das consequências das ações, sem requerer mapas dos domínios.

Q-Learning (ou Sarsa) pode servir de base para implementações mais sofisticadas, com o uso de sinais de elegibilidade (do Inglês, *eligibility traces*), por exemplo. Tais sinais, de acordo com a visão teórica, são uma ponte de métodos de Diferença Temporal (TD, sigla inglesa) para os de Monte Carlo, conforme mostra a Figura 2.5 [Sutton e Barto 1998]. A forma básica de métodos TD possui sinal de eligibilidade, denotado por λ, igual a zero, conhecido por TD(0), como mostra o diagrama TD de um-passo (primeiro diagrama de *backup* abaixo). Os métodos que estendem a diferença temporal por n passos são chamados métodos de Diferença Temporal n-passos (ou TD n-passos).

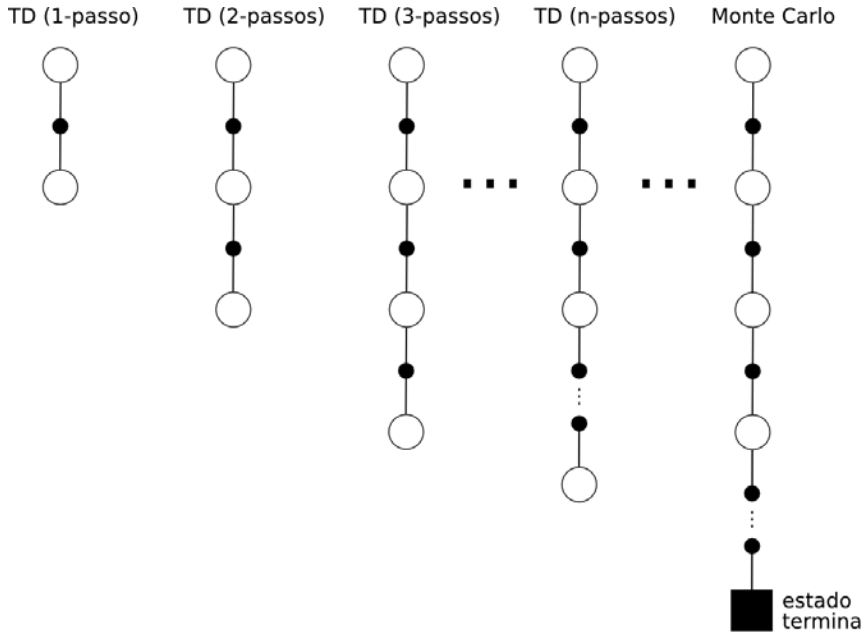

Figura 2.5: *Espectro variando de backups de métodos TD de um-passo até backups de métodos de Monte Carlo (Fonte: [Sutton e Barto 1998]).*

Diagramas de *backup* mostram todas as transições de estados que contribuem para atualização das funções-valor. Enquanto os diagramas de Programação Dinâmica (veja Figura 2.2) apresentam todas as transições

possíveis em um-passo, os diagramas de Monte Carlo mostram somente as transições amostradas até o fim do episódio, representando todo o percurso. Apoiada por recursos como sinais de elegibilidade ou Redes Neurais Artificiais para aproximação de funções [Saeb, Weber e Triesch 2009], a Aprendizagem por Diferença Temporal obtém as melhores características dos métodos de Monte Carlo e de Programação Dinâmica e, por consequência, alcança bons resultados.

2.2 Análise Experimental do Comportamento

A Análise Experimental do Comportamento (AEC) é uma disciplina científica surgida, em 1938, no contexto da psicologia com o objetivo geral de descrever e explicar as interações entre o organismo/indivíduo e o ambiente [Catania 1999]. Segundo [Skinner 2003], a AEC é uma ciência elementar do comportamento, onde é possível ser científico sem os recursos matemáticos da Ciência relacionados a observações quantitativas, mas não necessariamente prescindindo destes. Skinner ressalta que o comportamento é um objeto de estudo muito complexo, porém acessível e observável tanto em ambiente controlado quanto natural.

De forma geral, analistas do comportamento (filosoficamente orientados pelo behaviorismo) seguem o método indutivo para conduzir suas pesquisas. A partir de fatos (evidências), *behavioristas* validam dados empiricamente a fim de transformá-los em proposições gerais. Mesmo diante da imensa quantidade de variáveis, a análise experimental tem identificado regularidades no comportamento [Skinner 1950] [Baum 2005] [Neto 2002].

Na Educação, muitos procedimentos para a promoção da aprendizagem resultam de estudos comportamentalistas. Esta seção da Fundamentação Teórica apresenta a Análise Experimental do Comportamento com foco na história e em conceitos essenciais, como comportamento operante, condicionamento e reforço. Por fim, explica-se o Princípio de Premack, fundamento da estrutura do sistema proposto.

2.2.1 Breve Histórico

Nos primeiros anos do Século XX, o fisiólogo russo Ivan Pavlov investigou o comportamento animal em termos de associações entre estímulos e reflexos correlatos. Experimentos com cães ofereceram evidências sobre reflexos condicionados, aprendidos. Pavlov condicionou a secreção salivar de um cão a um tom, por emparelhamento temporal do tom com alimento. O cão respondeu à associação dos estímulos (tom e alimento), de modo que o estímulo incondicionado (alimento) legou suas funções ao estímulo originalmente neutro (tom) que, com função eliciadora condicionada, passou a produzir a mesma resposta (salivação) [Skinner 2003].

Em 1913, John Broadus Watson lançou o manifesto “A Psicologia tal como um behaviorista a vê”, destacando a Psicologia como estudo do comportamento [Skinner 2006]. Baseado nos experimentos de Pavlov, em 1916, Watson enfatizou a influência do ambiente na aprendizagem em seu discurso pre-

sidencial na Associação Americana Psicológica [Goulart 2007]. Desse modo, surgia formalmente o "*Behaviorismo*", como uma filosofia sobre o comportamento. Neste momento histórico, o entendimento conceitual sobre comportamento era o de ser este um conceito referente a respostas observáveis (manifestas) do organismo/indívíduo. Com a evolução do pensamento behaviorista, esta conceituação geral inicial foi (e tem sido) amplamente revisada, de modo a incluir eventos comportamentais não diretamente observáveis (por exemplo, o pensamento, a imaginação, a lembrança, etc.) [Skinner 2006] [Skinner 2003] [Catania 1999].

Contemporâneo de Watson, o psicólogo americano Edward Thorndike apresenta, em 1911, um princípio da psicologia da aprendizagem chamado "Lei do Efeito". Esta lei tratava de relações entre o comportamento e as consequências que este pode produzir no ambiente e, especificamente, sobre como tais consequências podem afetar o comportamento que as produziu. Diferentemente dos reflexos condicionados, ações espontâneas podem produzir consequências ambientais agradáveis, aumentando a chance de ocorrência do comportamento em situação semelhante. De outro modo, consequências ambientais desagradáveis implicam na diminuição da chance de ocorrência do comportamento.

A partir de seus experimentos, Thorndike demonstrou curvas de aprendizagem de respostas instrumentais (por exemplo, o acionamento de uma alavanca por um gato) como função de consequências viabilizadas pelas respostas (por exemplo, acesso a alimento). As curvas descreviam a diminuição gradual do tempo que decorria entre o animal ter a oportunidade de acionar a alavanca e a ação ocorrer (latência da resposta), na medida em que sucessivas oportunidades (tentativas) ocorriam. Além disto, o padrão do responder a alavanca sugeria, com o tempo, direcionalidade.

No final da década de 30, a pesquisa psicológica estava centrada no estudo da aprendizagem, especialmente de resposta tais como o percorrer um labirinto. Crítico da esterilidade das investigações dessa época, o psicólogo americano Burrhus F. Skinner publicou, em 1938, "O Comportamento dos Organismos" [Goulart 2007]. Skinner, sobre os ombros de Pavlov e Thorndike, reconheceu a existência de duas classes de comportamento: respondente (ou reflexo) e operante. Enquanto no comportamento respondente o estímulo elicia a resposta, no comportamento operante a resposta emitida age no ambiente e sofre os efeitos das suas consequências [Holland e Skinner 1975].

Os desenvolvimentos científicos da Análise Experimental do Comportamento proveram a Skinner os fundamentos para a formulação da sua versão

da filosofia behaviorista [Skinner 2006]: o *Behaviorismo* Radical. Em seu livro "Ciência e Comportamento Humano" [Skinner 2003], publicado em 1953, Skinner justifica e reforça a idéia da Análise Experimental do Comportamento como uma legítima ciência do comportamento, caracterizada pela observação empírica de ações sob condições detectáveis de controle ambiental e, consequentemente, viabilizadora, para a Psicologia, dos grandes objetivos de todas as ciências: a descrição, predição e controle dos seus objetos.

Analistas do comportamento, desde dos históricos Pavlov e Thorndike até Skinner, Keller, Holland, Ferster, estes e outros, desenvolveram as bases teóricas para pesquisas comportamentais em diversas áreas, inclusive na Educação [Schultz 2005].

As ideias de Skinner sobre os processos de aprendizagem fundamentaram-se significativamente naquelas de Pavlov e Thorndike, no entanto a concepção skinneriana recebeu várias outras influências, sobretudo filosóficas, o que o levou a tratar o comportamento de forma funcional, objetiva e operacional, nos moldes do modelo das ciências naturais. Para Skinner, três fatores devem ser considerados quanto se pretende descrever, explicar e predizer o comportamento: 1) o próprio comportamento (manifesto ou encoberto, tecnicamente descrito como "resposta"), 2) as condições ambientais antecedentes à ocorrência do comportamento (tecnicamente denominadas "estímulos discriminativos") e 3) os eventos ambientais consequentes à ocorrência do comportamento (eventos estes que podem ter funções reforçadoras ou punitivas). A interrelação dinâmica destes três fatores define o conceito de contingência (no caso tríplice), que é a principal unidade de análise do comportamento na elaboração skinneriana.

Segundo Skinner [Skinner 2006] [Skinner 1969] [Catania 1999], análises experimentais, quasi-experimentais e naturalíticas das interações dos organismos/indivíduos com o ambiente, tendo o conceito de contingência como ferramenta conceitual, estabelecem as condições para uma ciência do comportamento humano, com potencial para revelar as leis que governam interações de interesse, em todos os níveis.

2.2.2 Comportamento Operante

Pesquisas em ambientes controlados evidenciam respostas emitidas espontaneamente, sem influência aparente de estímulos eliciadores[2]. Nos ca-

[2]Estímulos eliciadores são aqueles que provocam invariavelmente respostas, como acontece na relação respondente.

sos das emissões ditas espontâneas, o comportamento age ou opera sobre o ambiente e, portanto, recebe o nome de **comportamento operante**. Se a resposta for imediatamente seguida por um **estímulo reforçador**, termo técnico para recompensa, ela será emitida mais frequentemente em circuntâncias similares. Situações de reforçamento aumentam a frequência do comportamento operante [Holland e Skinner 1975] [Skinner 2003]. Em relações deste tipo, a retirada do estímulo reforçador implica em diminuição gradual do comportamento que o produzia, processo tecnicamente denominado "**extinção**".

Treinadores de circo, por exemplo, utilizam, em treinamento dos animais circenses, comida para recompensar ações que desejam ver instaladas nos animais. Segundo [Holland e Skinner 1975], diz-se que o comportamento é reforçado pela comida. Reforçar repetidas vezes um mesmo comportamento quando este ocorre, constitui um processo chamado **condicionamento**.

Reforços podem ser **incondicionados** (ou primários) ou **condicionados** (ou secundários). Os primeiros promovem o aumento da probabilidade de ocorrência do comportamento independentemente de uma história de aprendizagem para o estabelecimento desta função (por exemplo, alimento, água, orgasmo, remoção de eventos aversivos, etc.). Os segundos, condicionados, adquirem a função de reforço por emparelhamento, ao longo da história de vida, com os reforços incondicionados (por exemplo, a atenção, o dinheiro, boas notas, elogios, etc.). Reforços condicionados generalizados são eventos que funcionam como reforço independentemente do estado de privação em que o organismo pode se encontrar, sendo o dinheiro um dos principais exemplos.

Situações experimentais clássicas ilustram os conceitos apresentados. Numa delas, o comportamento de um pombo privado de alimento dentro de uma caixa fechada foi observado, conforme mostra Figura 2.6. Toda vez que o pombo bica o disco, automaticamente recebe comida pelo alimentador. Sob estas condições, a comida funciona como reforçador, pois aumenta a frequência de ocorrência da resposta de bicar o disco. Após a aprendizagem, caso o pombo emita respostas e não seja reforçado (procedimento de extinção), o comportamento de bicar o disco se extingue, ou seja, volta à baixa frequência de antes do condicionamento. Fazendo uma distinção importante, ao contrário da extinção, o **esquecimento** ocorre quando as respostas não são emitidas, visto não haver suporte ambiental para tanto [Holland e Skinner 1975].

Ainda em experimentos com pombos, alguns eventos são condicionados para reforçar o comportamento. Juntamente com a comida, o alimentador apresenta um ruído e ilumina o alimento servido como consequência do "bicar o disco". O som e a luz ficam evidentes enquanto o alimento estiver disponí-

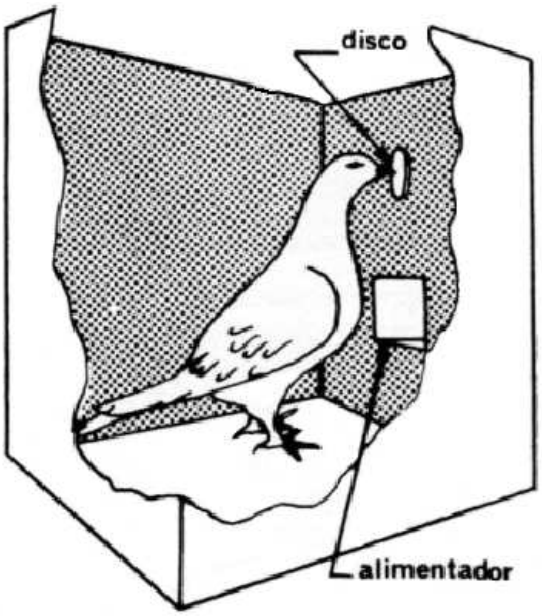

Figura 2.6: *Pombo em situação experimental típica (Fonte: [Holland e Skinner 1975]).*

vel para o pombo. Assim, após algum tempo, esses estímulos sonoro e visual adquirem a função de reforçadores (secundários).

Segundo [Skinner 2003], em muitas áreas, como na educação, na indústria, na psicoterapia, profissionais aplicam técnicas para criar reforçadores condicionados apropriados, a fim de atingir níveis melhores de respostas. Em termos práticos, a alteração dos eventos que definem as contingências pode incentivar o aluno a estudar, os empregados a serem assíduos e competentes, os pacientes a comportarem-se adequadamente.

Para obter comportamento operante socialmente desejável, respostas indesejáveis podem ser eliminadas por extinção ou pelo reforço de respostas incompatíveis. Um professor, ao dispensar uma turma desobediente, pode reforçar a indisciplina. De forma oposta, se o professor só liberar mais cedo os alunos quando se comportarem bem, o reforço da resposta incompatível pode eliminar o mau comportamento.

Em condições naturais ou controladas, a frequência do comportamento operante aumenta por meio de **reforçamento**, tanto **positivo** quanto **negativo**, ou seja, pela apresentação ou remoção dos estímulos reforçadores [Skinner 2003]. O reforço positivo, também chamado *recompensa*, constitui uma consequência apetitiva e o reforço negativo, *alívio*, implica remover uma consequência aversiva [Catania 1999].

Diferentemente do reforçamento, a **punição** enfraquece o comportamento, pois os seus efeitos são supressores. Existem dois tipos de punição, por

apresentação e por remoção, assim como no reforçamento. Num tipo introduz-se uma consequência negativa, *castigo* e, no outro, a consequência positiva é removida, ou seja, aplica-se uma *penalidade*.

Em síntese, a Análise do Comportamento investiga variáveis dependentes, como a taxa de aquisição de respostas, a partir da alteração de valores das variáveis independentes, como o reforçamento (e seus parâmetros). O pesquisador observa o comportamento e mede, quantifica, a rapidez para um comportamento se estabelecer (taxa de aquisição), a quantidade de respostas emitidas após certo período (taxa de resposta) e o tempo para eliminar respostas com a ausência de reforçamento (taxa de extinção). Porém essas variáveis podem ser complexificadas significativamente por tipos e esquemas de reforçamento [Catania 1999].

Os **esquemas** de **reforçamento** indicam quais, quando e como usar reforçadores. Afinal um mesmo reforçador pode reforçar numa circuntância e ser indiferente em outra. Definir que reforçador usar é um ponto essencial do reforçamento. Para saber quando e como reforçar, escolhe-se um esquema de reforçamento conforme modelo característico: contínuo, intermitente ou combinado. No esquema contínuo, ocorre o reforço de toda e qualquer resposta correta. No intermitente ou parcial, o reforçamento ocorre espaçado no tempo (intervalo) ou em função do número de respostas emitidas (razão).

Em esquemas intermitentes de razão fixa, a resposta n de n respostas é reforçada, ou seja, num esquema RF5 (Razão Fixa 5), a quinta resposta produz reforço e as respostas 1, 2, 3 e 4, não. Nos esquemas de razão variável, uma resposta em torno de um número n médio de respostas é reforçada, ou seja, num esquema RV3 (Razão Variável 3), uma resposta é reforçada a cada 3 respostas emitidas em média. Por exemplo, numa sequência de 5 reforços, um pode vir para a 2ª resposta, o próximo para a 6ª (após a 2ª), o próximo para a 1ª (após a 6ª), o próximo para 3ª (após a 1ª), e o próximo, novamente, para a 3ª resposta (após a 3ª anterior).

Nos casos de reforçamento intermitente de intervalo fixo, a primeira resposta emitida após transcorrido um intervalo de tempo fixo é reforçada, ou seja, um IF1' (Intervalo Fixo 1 minuto) especifica que apenas a resposta emitida após o transcurso de um minuto desde o último reforço será, caso emitida, reforçada. Nos esquemas de intervalo variável, a primeira resposta emitida após transcorrido um intervalo médio de tempo é reforçada, ou seja, um IV1' (Intervalo Variável 1 minuto) especifica que, em média, a cada um minuto o reforçador pode ser obtido caso ocorra uma resposta. Por exemplo, numa seqüência de 5 reforços, o primeiro pode ser obtido após 40", o segundo

após 120", o terceiro após 60", o quarto após 35" e o quinto após 45".

No esquema de reforçamento contínuo, a aquisição e a extinção do responder são rápidas. Nos esquemas intermitentes, onde pequenos períodos de extinção ficam entremeados às ocorrências do reforço, a aprendizagem é mais lenta, dura mais tempo, sendo mais resistente à extinção.

A Análise Experimental do Comportamento apresenta bons resultados em muitos ambientes, em escolas, empresas e instituições de saúde. Embora sejam comuns práticas sociais leigas baseadas em contingências aversivas, existem evidências empíricas sobre os benefícios do uso de contingências positivas por comunidades de analistas comportamentais. Como exemplo, a instrução programada, desenvolvida por Skinner, constitui uma aplicação educacional com preparo minucioso do material. O aluno avança para novos conteúdos no seu próprio ritmo e as notas indicam quanto ele sabe da matéria, ao invés de serem taxativas[3] [Skinner 1972].

Diante da proposta da instrução programada, destacou-se o uso de máquinas para interagir com o estutante. A evolução tecnológica fez dos computadores as máquinas de ensinar de hoje. Além de apresentar o conteúdo para o aluno com *feedbacks* imediatos e interatividade, eles dispõem de recursos para atender o estudante de forma diferenciada, por meio de técnicas de Inteligência Computacional. Ao longo de duas décadas, investigações apontam para ganhos significativos no desempenho do aluno quando este aprende com sistemas tutores inteligentes, os quais buscam adaptar-se às necessidades do aprendiz.

2.2.3 Princípio de Premack

Premack, em experimentos com macacos *Cebus* e crianças, verificou que atividades com elevada frequência podem funcionar como reforçadores para atividades de baixa frequência. Tal constatação foi formulada como o princípio de Premack [Premack 1959]. Ele identificou atividades como reforçadoras, além dos estímulos, e assumiu a relatividade do reforçamento, sendo que respostas naturalmente mais frequentes para o organismo/indivíduo reforçam respostas menos frequentes.

Há diversas aplicações da Análise Experimental do Comportamento, com autistas, esquizofrênicos, estudantes, funcionários, entre outros. Referentes ao princípio de Premack, alguns estudos são listados abaixo.

[3]De modo geral, as escolas usam as notas baixas como consequências aversivas (punição), assim o aluno estuda para não tirar nota baixa, sem desejar aprender de fato.

- Área educacional: “Um tempo para aprender, um tempo para brincar: Princípio de Premack aplicado em sala de aula” [Geiger 1996] e “Tempo livre como um reforçador no controle do comportamento em sala de aula” [Osborne 1969];
- Área organizacional: “Princípio de Premack aplicado para melhorar a qualidade dos serviços feitos pelos empregados” [Welsh, Bernstein e Luthans 1993];
- Área da saúde: “Aplicação do Princípio de Premack para controle comportamental de esquizofrênicos extremamente inativos” [Mitchell e Stoffelmayr 1973].

Na área educacional, Osborne (1969) contingenciou tempo livre das atividades escolares como consequência para a permanência nos assentos na sala de aula (resposta operante), numa turma de seis alunos surdos. Como resultado, a frequência do comportamento dos alunos de ficarem sentados aumentou, mesmo após a suspensão da contingência (comportamento observado por seis semanas) [Osborne 1969]. Geiger (1996) aplicou o princípio de Premack em turmas de sétima e oitava séries. Caso o grupo/turma terminasse as atividades de 5 a 10 minutos antes do término da aula, era liberado para brincar no *playground*. Evidenciou-se maior disciplina e maior aproveitamento escolar para o grupo experimental, comparado ao grupo controle (aula tradicional) [Geiger 1996].

Atualmente, as Tecnologias da Informação e Comunicação (TICs) são poderosos meios de entretenimento. Naturalmente e com alta frequência, crianças e adolescentes utilizam o computador para realizar algo de seu interesse [Subrahmanyam et al. 2001], como jogar, assistir a vídeos, comunicar-se via Internet. Nesse contexto, o presente estudo contingencia tempo livre, quando o aprendiz podia jogar, ouvir música (atividades comumente mais frequentes), como consequência às atividades de estudo no tutor (atividades comumente menos frequentes). Sendo assim, aplicou-se o princípio de Premack como fundamento para o sistema proposto, detalhado no próximo capítulo.

CAPÍTULO 3

Sistema Proposto

Esta pesquisa propõe o controle inteligente de tempo livre em tutoria multissessão, onde as etapas (sessões) da tutoria são separadas por tempo livre, pausas. Baseado no desempenho do aluno, o sistema proposto utiliza Aprendizagem por Reforço para controlar a duração das pausas. Este capítulo apresenta a estrutura do sistema tutor e o controle inteligente de tempo livre.

3.1 Estrutura do Sistema Tutor Inteligente

Em Sistemas Tutores Inteligentes, o cenário de tutoria interfere no projeto instrucional e na modelagem da técnica inteligente. Nesta pesquisa, a estrutura do sistema tutor possui **tempo livre** contingenciado ao tempo em **tutoria**. A pausa entre sessões de tutoria caracteriza o cenário proposto.

3.1.1 Tempo Livre

Após certo período de estudo, a capacidade do aluno de reter conhecimento diminui. Naturalmente, o aprendiz deseja um momento de descanso. **Tempo livre** é a **pausa necessária** entre sessões de tutoria. Como reforço para a tutoria, o tempo livre pode ser condicionado adequadamente e melhorar o desempenho do estudante na atividade de aprendizagem [Geiger 1996] [Osborne 1969].

Baseado na Análise Experimental do Comportamento, Premack [Premack 1959], em seu Princípio, postula que atividades mais prováveis reforçam atividades menos prováveis. No sistema proposto, o estudante, após passar por uma sessão de tutoria, conquista liberdade para agir por certo tempo. Em síntese, existem duas situações de interação: na tutoria (estudo) e fora da tutoria (tempo livre). Atividades no tempo livre, mais prováveis, reforçam a dedicação no tempo de estudo, menos provável em condições normais.

A **duração** do tempo livre pode ser **predefinida** ou **controlada** durante a tutoria, como exemplifica a Figura 3.1. A pausa definida *a priori*, constante ou não, mantém-se a mesma para todos os alunos. Ao contrário, quando há controle do tempo livre, o sistema tutor determina o valor da pausa no fim de cada sessão de tutoria. O novo tempo livre pode ser calculado em função do tempo livre anterior t_{i-1} combinado a uma ação a_i escolhida pelo aluno ou pelo próprio sistema.

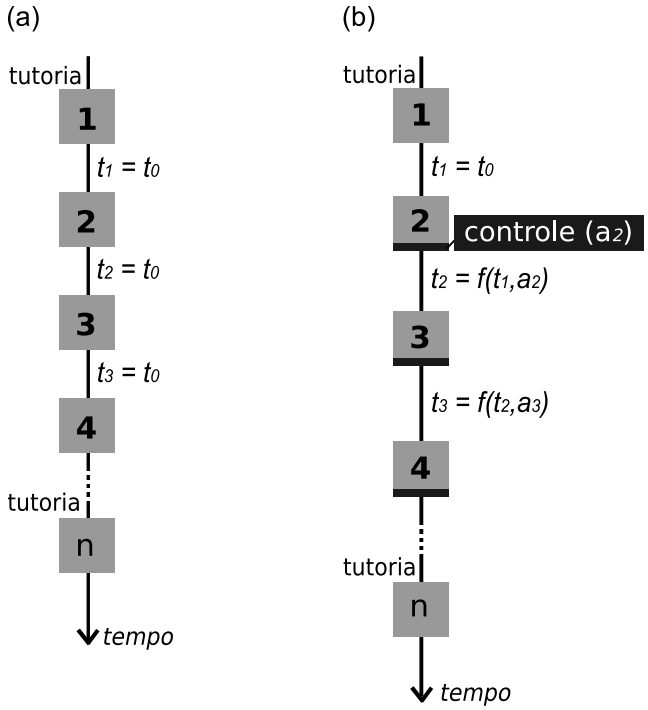

Figura 3.1: *Tutoria com tempo livre (a) predefinido versus (b) controlado.*

O controle inteligente proposto utiliza Aprendizagem por Reforço para determinar a duração adequada da pausa para o aprendiz. De forma detalhada, a Seção 3.2 deste capítulo responde à pergunta: "Como controlar o tempo livre em Sistemas Tutores Inteligentes?".

3.1.2 Tutoria

Softwares educacionais auxiliam o estudante no processo de aquisição de conhecimento, apresentando conteúdo minuciosamente programado [Skinner 1972], elaborado [Wilson e Cole 1992]. A tutoria é o mecanismo de ensino, ou seja, a experiência de interação entre o aluno e o sistema tutor inteligente. Para compor a tutoria, propõem-se as seguintes etapas: 1) Abordagem Inicial, 2) Curso (Conjunto de Módulos) e 3) Abordagem Final. Primeiro,

o sistema tutor estabelece comunicação com o aluno: Abordagem Inicial (apresentação, questionários, etc). Em seguida, o Curso explica o conteúdo ao estudante. Por fim, aplica-se alguma avaliação e encerra a tutoria (Abordagem Final), como mostra a Figura 3.2.

O Curso, etapa intermediária, é um conjunto de módulos, compostos por vídeo-aula, exercício, sugestão prática, tempo livre e exercício de revisão. Como proposta deste trabalho, as vídeos-aulas devem ser curtas, com o conteúdo apresentado de forma esquemática (palavras-chave e imagens). Em sincronia com a parte visual, a narração explica o assunto numa linguagem apropriada ao aluno.

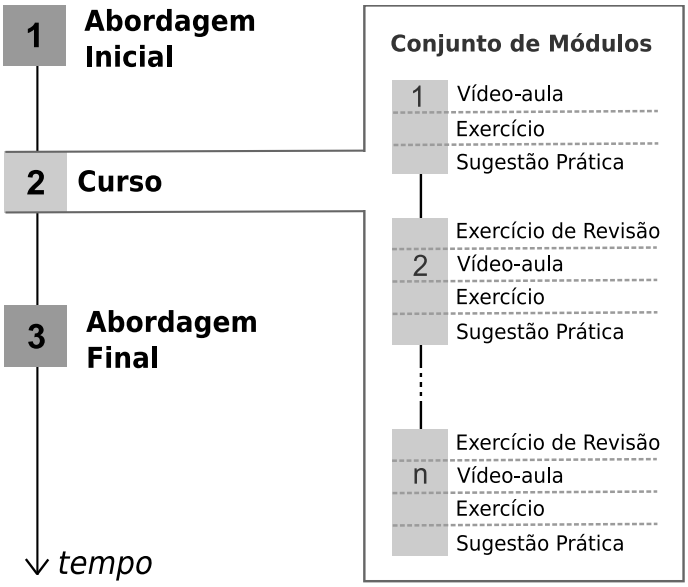

Figura 3.2: *Etapas da tutoria.*

Após a vídeo-aula, o aprendiz responde a um exercício, avaliação conceitual. Imediatamente depois do tempo livre, o sistema tutor apresenta outro exercício, de revisão, com caráter procedural ou teórico. Os exercícios possuem os seguintes tipos de resposta: 1) Correta, 2) Semelhante à Correta e 3) Incorreta. O estudante pode responder à pergunta com exatidão (tipo 1) ou de forma vaga, genérica (tipo 2) ou errar (tipo 3). Em cada pergunta, o sistema tutor pede ao aluno para escolher a melhor alternativa, ou seja, a resposta correta (precisa, exata). Para cada tipo de resposta escolhida, o estudante recebe uma mensagem diferente de *feedback*. A Figura 3.3 mostra um exemplo dos três tipos de resposta, onde a alternativa “a” é a resposta Correta, a alternativa “b” é a resposta Semelhante à Correta e os itens “c” e “d” são respostas Incorretas.

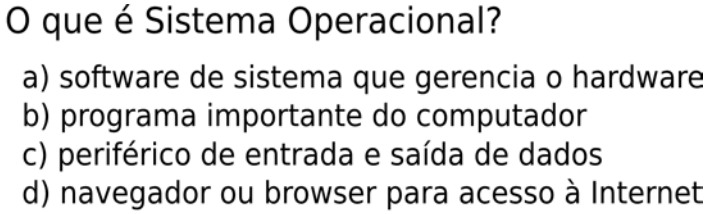

Figura 3.3: *Exemplo de tipos de respostas.*

A partir dos estudos de Pressey, avaliar pôde ser mais uma forma de ensinar [Skinner 1972] [Candau 1969]. No sistema proposto, os exercícios promovem o aprendizado, visto que o estudante só prossegue depois de acertar. A cada resposta dada, o aluno recebe reforços positivos e negativos. Se for a alternativa correta, o sistema retira as outras alternativas (reforço negativo) e destaca a resposta correta (reforço positivo). Caso o aprendiz escolha uma alternativa diferente da correta, o sistema tutor elimina o item da tela (reforço negativo) e pede para tentar outra alternativa (reforço positivo).

Cada módulo (ou ciclo) do curso possui as seguintes fases: 1) Vídeo-aula, 2) Exercício, 3) Sugestão Prática, 4) Tempo Livre e 5) Exercício de Revisão (Figura 3.4). Além de ensinar a teoria e aplicar exercícios, o sistema tutor inteligente estimula o estudante a desempenhar alguma atividade no seu tempo livre, conforme sugestão dada.

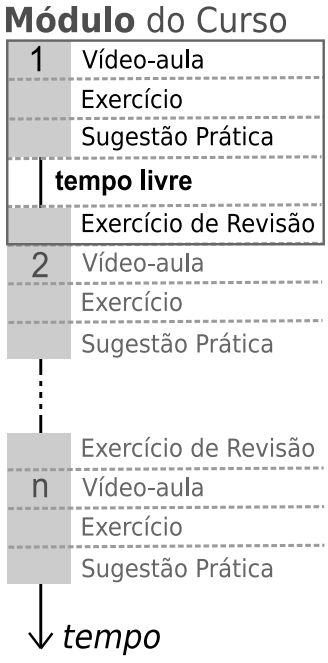

Figura 3.4: *Módulo (ou ciclo) do curso.*

Num curso de dança, por exemplo, o aluno pode praticar algum passe no seu tempo livre. No caso de um curso de piano, o aprendiz pode tocar um pouco e assim por diante. Após a pausa, o sistema tutor retorna com o exercício de revisão, uma questão sobre algum procedimento ou sobre a teoria do módulo.

Segundo [Thierauf 1995], as pessoas retém aproximadamente 25% do que ouvem, 45% do que ouvem e veem, e 70% do que ouvem, veem e fazem. Portanto o ensino com vídeo-aulas e tempo livre para praticar possibilita maior retenção de conhecimento. A participação ativa do aluno na aprendizagem contribui para que ele aprenda de forma duradoura [Schank 1994].

3.2 Controle Inteligente do Tempo Livre

O controle inteligente desta pesquisa utiliza Aprendizagem por Reforço para adaptar o tempo livre ao perfil do aluno. Esta técnica estrutura-se em estado-ação-recompensa, ou seja, o agente escolhe uma ação a no estado s e, como resposta do ambiente, recebe uma recompensa r.

No sistema proposto, cada estado s refere-se a um módulo do curso (passo do episódio). Para passar de um estado para outro, aplica-se uma ação a do conjunto de ações A, $\{diminuir, manter, aumentar\}$ a duração da pausa. Por fim, baseado no desempenho nos exercícios (notas $N1$ e $N2$), calcula-se a recompensa r, como mostra a Figura 3.5.

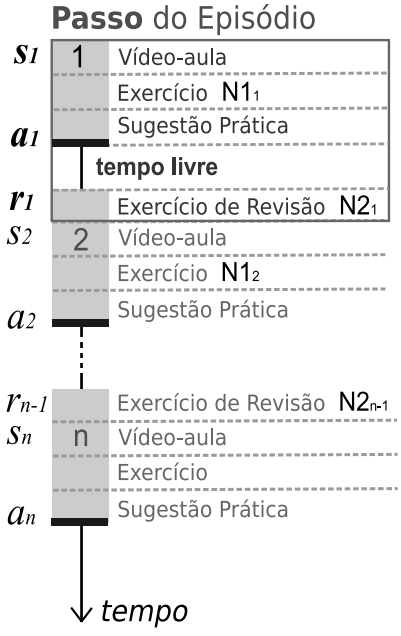

Figura 3.5: *Estrutura do sistema proposto.*

Em cada estado s, o agente 1) seleciona uma ação a a partir da política $\pi_t(s,a)$, 2) determina r pela função-recompensa e 3) atualiza o valor da ação $Q(s,a)$, conforme algoritmo da Figura 3.6. As seções abaixo explicam cada etapa detalhadamente.

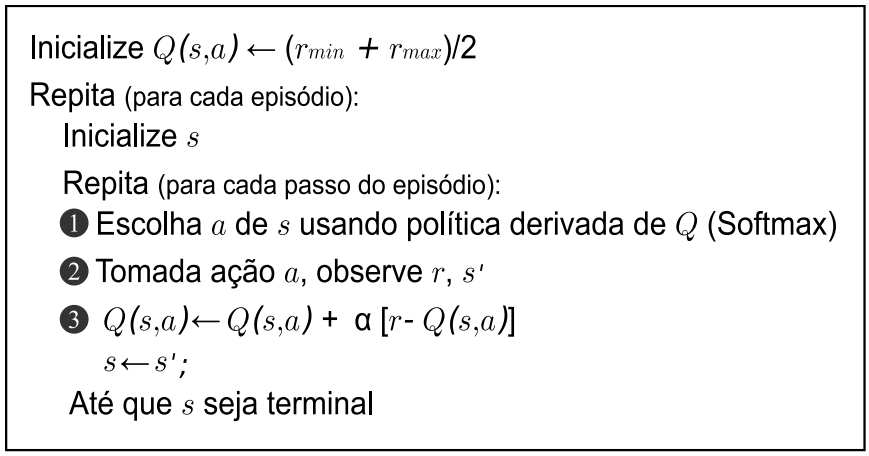

Figura 3.6: *Algoritmo do sistema proposto.*

Os passos 1, 2 e 3 podem ser contextualizados em relação aos principais componentes de um sistema tutor inteligente (ver Figura 3.7). O passo 1 corresponde ao módulo pedagógico, onde a política escolhe uma ação no estado atual, a estratégia de ensino apropriada para o STI aplicar. Tal escolha deriva de Q (valor da ação) para cada aluno, ou seja, deriva do modelo do estudante. No passo 2, a estratégia, aplicada ao conhecimento do domínio, gera uma recompensa, avaliação de desempenho, e leva a um novo estado, módulo, a ser apresentado para o aprendiz usando o modelo de comunicação. Assim que o aluno responde aos exercícios, o valor Q (modelo do estudante) é atualizado (passo 3) e o ciclo repete.

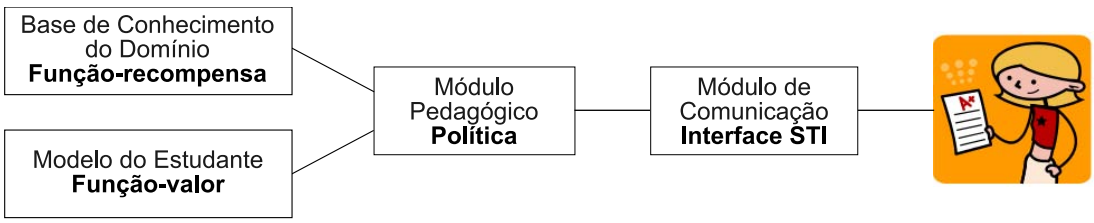

Figura 3.7: *Principais componentes do STI proposto.*

3.2.1 Política

A política π define, através do **método Softmax**, as probabilidades de ocorrência de cada ação. Este método ordena as ações de acordo com os seus

valores, calibrado pelo parâmetro τ (temperatura) da distribuição de Gibbs, conforme a Equação 2-8. Para valores altos da temperatura, as ações tornam-se equiprováveis. À medida que a temperatura tende a zero, o método Softmax assemelha-se ao método Greedy, onde escolhe-se a ação de maior valor.

O **ajuste da temperatura** é um ponto crucial para convergência. Portanto este trabalho propõe caimento linear da temperatura com razões distintas para dois momentos da aprendizagem computacional: 1) exploração extensiva e 2) exploração intensiva. A Figura 3.8 mostra a proposta de ajuste da temperatura, com as razões q_1 e q_2 calculadas pelas Equações 3-1 e 3-2.

$$q_1 = \frac{\tau_I - \tau_0}{s_I - s_0} \tag{3-1}$$

$$q_2 = \frac{\tau_F - \tau_I}{s_F - s_I} \tag{3-2}$$

Os valores para τ_0 (temperatura inicial), τ_I (temperatura intermediária) e τ_F (temperatura final) são obtidos por meio de simulações, de acordo com o cenário em questão. As temperaturas inicial, intermediária e final correspondem aos valores ajustados para os estados inicial s_0 (referente ao primeiro módulo), intermediário s_I e final s_F (referente ao último módulo), respectivamente. Ainda nas simulações, define-se o estado intermediário, ou seja, o ponto a partir do qual o agente inteligente aproveita o conhecimento (exploração intensiva) adquirido no treinamento (exploração extensiva).

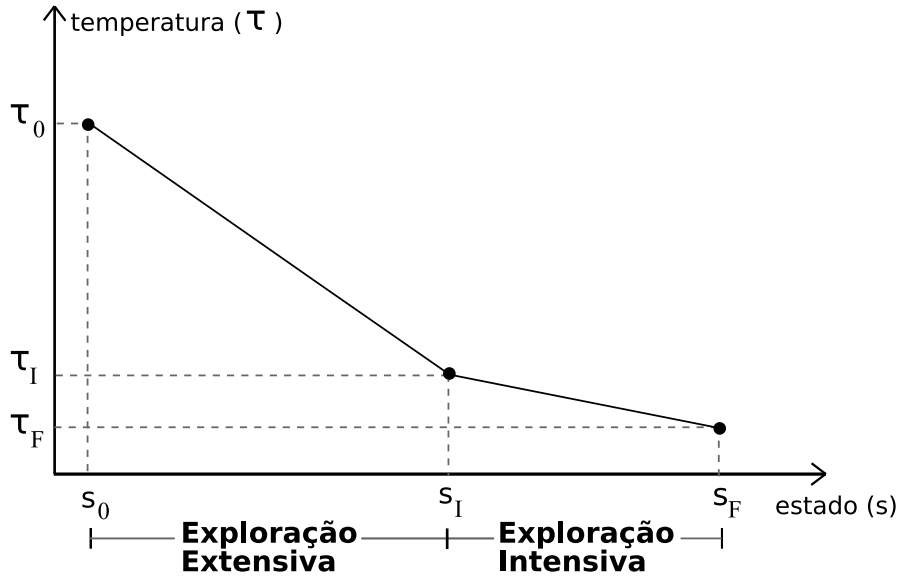

Figura 3.8: *Ajuste da temperatura com razões distintas para o caimento.*

Segundo a Análise Experimental do Comportamento, o elemento reforçador para o estudante diz respeito ao controle do tempo livre, à mudança ou não da duração da pausa como consequência do seu desempenho nos exercícios. A variação da pausa tem por objetivo atingir o tempo adequado ao aluno, em que há maior retenção de conhecimento. Ao contrário, na Aprendizagem por Reforço, o reforço (recompensa) para o agente é o desempenho do aluno, as notas nos exercícios. Essa diferenciação, embora complexa, contextualiza o sistema proposto em Psicologia e em Engenharia (ver tabela abaixo).

Tabela 3.1: *Contextualização do sistema proposto em Psicologia e em Engenharia.*

Elemento \ Área	**Psicologia**	**Engenharia**
Disciplina	Análise do Comportamento	Aprendizagem por Reforço
Sujeito	Estudante	Agente
Ação	Responder aos exercícios	Controlar o tempo livre
Reforço	Controle do tempo livre	Nota nos exercícios

A fim de verificar o comportamento do sistema proposto, a modelagem do problema da Aprendizagem por Reforço pressupõe diferentes **perfis** de aluno: **teórico**, **equilibrado** e **pragmático**. Em simulações para ajuste da temperatura, o desempenho do aluno teórico melhora quando o tempo livre diminui, do aluno equilibrado, quando tempo livre mantém-se e, do pragmático, quando aumenta. Vale ressaltar que os tipos de estudante foram definidos para promover a convergência da Aprendizagem por Reforço (para uma ação), a princípio por meio de simulações e *a posteriori*, colocados à prova, através da coleta de dados com humanos.

3.2.2 Função-recompensa

Para encontrar a melhor ação, o agente modifica a duração da pausa a partir do tempo livre inicial t_0. De forma extensiva, o agente experimenta as ações disponíveis. Depois, com base no conhecimento adquirido, ele prioriza a ação de maior valor, ou seja, a mais recompensada (exploração intensiva).

Na Figura 3.5, $N1$ e $N2$ são as notas do aluno em cada exercício, após vídeo-aula e após tempo livre (exercício de revisão), respectivamente. A função-recompensa (ver a Equação 3-3) determina o valor da recompensa r. Como exemplo desta função, r pode ser a média ponderada de $N1$ e $N2$, como mostra a Equação 3-4, onde $L > K$ para enfatizar o momento posterior à ação a aplicada.

$$r = f(N1, N2) \tag{3-3}$$

$$r = \frac{K.N1 + L.N2}{K + L} \tag{3-4}$$

Tanto $N1$ quanto $N2$ são valores numéricos para o desempenho do aluno, de 0 a 1, ou seja, de nenhum a total conhecimento da resposta. Acertar o exercício é a condição para seguir adiante na tutoria, logo $N1$ e $N2$ resultam da quantidade de tentativas relacionadas aos tipos de resposta. Se o aluno acerta na primeira tentativa, ele conseguirá nota máxima, 1. Se o acerto ocorrer somente na segunda, terceira ou n-ésima tentativa, a nota diminui devido à perda por tentativa p_j (como mostra a Equação 3-5). No pior caso, a perda total (somatório de perda por tentativa) dividida pela perda de referência é igual a 1, portanto a nota terá o menor valor possível, 0 (zero).

$$Nota = 1 - \frac{\sum_{j=1}^{ta} p_j}{p_R} \tag{3-5}$$

Nas equações acima, p_R é a perda de referência (valor da perda total no pior caso) e ta é a tentativa do acerto. Para cada tipo de resposta (Correta, Semelhante à Correta, Incorreta) existe um peso específico no cálculo da perda, definido pela função $H(x)$. Por exemplo, se x for igual à resposta correta, então $H(x) = 0$, de forma que não haja perda neste caso. A Equação 3-6 mostra o cálculo da perda.

$$p_j = H(x).\frac{1}{T - j + 1} \tag{3-6}$$

Onde *T* é o total de alternativas, ou seja, o número máximo de tentativas até acertar o exercício. À medida que o estudante erra, as alternativas escolhidas desaparecem da tela e a perda por tentativa aumenta, devido à chance de acerto ser maior (menos opções de escolha).

3.2.3 Função-valor

Em métodos de Aprendizagem por Reforço, o agente atualiza o valor da ação $Q(s, a)$ utilizando a recompensa r. Segunto [Sutton e Barto 1998], valor ou funções-valor de ação (pares estado-ação) são características-chave desses métodos, pois o valor da ação armazena o conhecimento do agente sobre o ambiente. Para favorecer a exploração extensiva, o algoritmo proposto

inicializa $Q(s, a)$ com um valor intermediário de r, valor inicial otimizado, conforme a instrução a seguir.

$$Q(s, a) \leftarrow \frac{r_{min} + r_{max}}{2} \tag{3-7}$$

O cenário de tutoria apresenta uma sequência limitada de passos, caracterizando uma tarefa episódica. Portanto, ao final de cada passo (módulo) do episódio (curso), o valor da ação $Q(s, a)$ é atualizado de forma incremental, como mostra a instrução abaixo (ver também o algoritmo da Figura 3.6).

$$Q(s, a) \leftarrow Q(s, a) + \alpha[r - Q(s, a)] \tag{3-8}$$

O sistema proposto estima o valor das ações como a média das recompensas observadas, portanto o parâmetro α, conhecido como o tamanho do passo (do Inglês, *step-size*), é igual $\frac{1}{k}$, para k-ésima recompensa recebida para ação a.

CAPÍTULO 4

Experimento e Resultados

Este capítulo apresenta o experimento realizado, desde a modelagem matemática do algoritmo até à coleta de dados, e os resultados obtidos a partir da análise e interpretação dos dados. O experimento define as etapas da tutoria, os estados, ações e recompensas, bem como a função-valor da Aprendizagem por Reforço e a temperatura, parâmetro do método Softmax para seleção de ação.

Quanto aos resultados, as análises descritiva e inferencial dos dados permitiram a verificação da hipótese básica desta pesquisa: "Se utilizar controle inteligente do tempo livre, então aluno retém mais conhecimento". O grupo experimental (com controle inteligente do tempo livre) foi comparado ao grupo controle (onde a decisão pertence ao próprio estudante). Resultados mostram ganhos significativos e equivalentes na retenção de conhecimento. Contudo, alunos do grupo experimental perceberam melhor o tempo livre como componente da estratégia de ensino.

4.1 Experimento

O experimento implementa um sistema tutor inteligente para ensinar conceitos básicos de Microinformática no Linux. Quanto à Aprendizagem por Reforço, os estados referem-se aos quinze módulos do conteúdo e as ações podem alterar ou manter a duração do tempo livre. De acordo com as condições experimentais, simulações contribuíram no ajuste do parâmetro τ, temperatura, e do valor inicial das ações, a fim de balancear exploração extensiva e intensiva. Para determinar o valor de cada ação, o agente calcula a média das recompensas recebidas de forma incremental (função-valor). As recompensas resultam da medida do desempenho do aluno nos exercícios do curso (função-recompensa). Detalhadamente, as subseções a seguir explicam o experimento.

4.1.1 Etapas da Tutoria

As etapas da tutoria compreendem todo conteúdo do sistema tutor inteligente acrescido de tempo livre. Basicamente, uma sessão de tutoria possui um questionário ou um módulo do curso. Entre uma sessão e outra, há o tempo livre, uma fase especial da tutoria, pois está inserido no contexto de ensino. A Figura 4.1 representa as etapas do sistema tutor implementado: 1) Apresentação (do sistema), 2) Avaliação de Conhecimento Prévio (preteste), 3) Curso (Conjunto de Módulos), 4) Avaliação de Conhecimento Posterior (posteste), 5) Avaliação de Percepção da Experiência e 6) Encerramento.

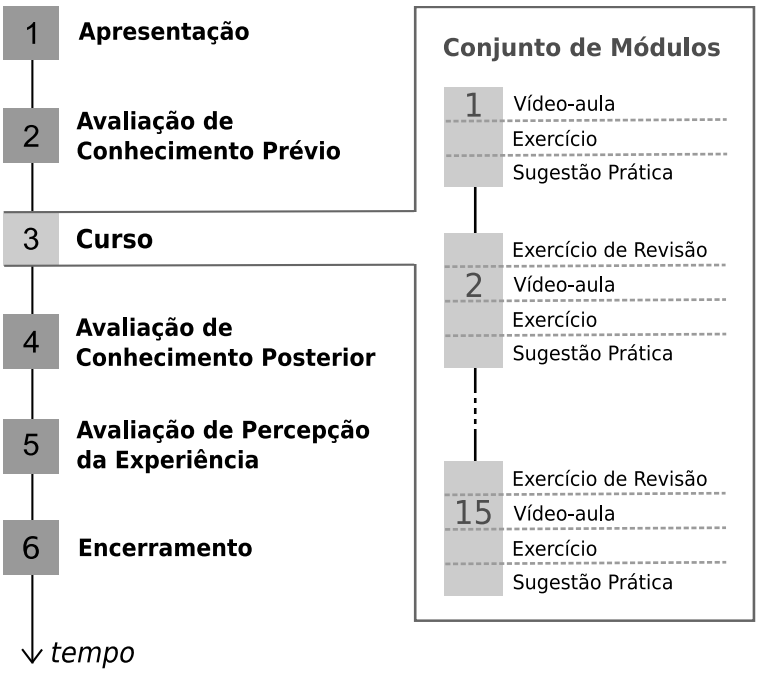

Figura 4.1: *Etapas da tutoria implementadas.*

A "Abordagem Inicial" do sistema proposto (Capítulo 3) corresponde às Etapas 1 e 2, Apresentação e Avaliação de Conhecimento Prévio. Enquanto a "Abordagem Final" subdivide-se nas etapas 4) Avaliação de Conhecimento Posterior, 5) Avaliação de Percepção da Experiência e 6) Encerramento.

Inicialmente, na primeira etapa, o estudante assiste a um tutorial em vídeo sobre o sistema tutor, contendo capturas de telas e explicações a respeito da interação com o sistema, além de esclarecer alguns conceitos, como o significado de "Tempo Livre". Na etapa 2, Avaliação de Conhecimento Prévio, o sistema tutor apresenta quinze questões sobre "Microinformática no Linux" (conteúdo do curso). Cada pergunta refere-se a um módulo do curso, apresentada na mesma ordem dos assuntos. Por exemplo, a Questão 1 refere-

se ao Módulo 1, “Introdução”, a Questão 2 refere-se ao Módulo 2, “Hardware”, e assim sucessivamente.

O Curso, principal parte da tutoria, é um conjunto de quinze módulos com atividades de ensino. O conteúdo do curso “Introdução à Microinformática no Linux” foi separado por assuntos: 1) Introdução, 2) Hardware, 3) Partições, 4) Software, 5) Sistema Operacional, 6) Linux, 7) Área de Trabalho, 8) Aplicativos, 9) Dados, 10) Diretórios, 11) Arquivos, 12) Permissões, 13) Compactadores, 14) Internet e 15) Conclusão. Os módulos abordam o tema em questão por meio de vídeo-aula, exercício, sugestão prática, tempo livre e exercício de revisão, nessa ordem.

A vídeo-aula, com três minutos de duração em média, apresenta o assunto de forma simples para o aluno, mostrando analogias ilustradas e esquemas com palavras-chave em destaque. Os exercícios são de múltipla escolha e devem ser respondidos corretamente, mesmo que após muitas tentativas. O sistema tutor aguarda o acerto do estudante para liberar a acesso à próxima etapa. Como explicado no Capítulo 3, existem três tipos de respostas, Correta (C), Semelhante à Correta (SC) e Incorreta (I). Dentre as quatro alternativas, uma é a resposta Correta, outra é a Semelhante à Correta e as duas restantes são Incorretas. Quanto à sugestão prática, o sistema apresenta imagens com procedimentos para o aluno executar no tempo livre, caso queira.

Após o Curso, na Avaliação de Conhecimento Posterior (Fase 4), o aluno responde às mesmas questões da Avaliação de Conhecimento Prévio, a fim de mensurar a retenção de conhecimento (comparando a nota final com a inicial). Em seguida, o aluno responde a seis questões, na Etapa 5, sobre a sua percepção quanto à experiência no sistema tutor. Para finalizar, o Encerramento (Fase 6) contém uma mensagem de agradecimento e duas perguntas, sendo a primeira “Você quer continuar no computador ou não?” e a segunda “Qual motivo?”. Assim, registra-se a permanência do aluno no computador (após o término da tutoria), a fim de fornecer informações interessantes, inclusive para Psicologia[1].

Entre as sessões de tutoria, o tempo livre possui valor predefinido entre as etapas gerais da tutoria e controlado durante o curso (ver a Figura 4.2). A partir da pesquisa-piloto e da estimativa da duração total da atividade, tanto de tutoria quanto de pausas, definiu-se o tamanho do tempo livre em

[1]As respostas dos alunos indicam a qualidade da interação Homem-Máquina. Investigações possíveis: 1) verificar a consistência do considerar o tempo livre insuficiente com o continuar no computador após a tutoria e 2) verificar correspondências entre dizer e fazer (comportamentos verbal e não-verbal) quanto à permanência ou não no computador.

quatro minutos ($t = 4min$) após a primeira e a segunda etapas e em um minuto ($t = 1min$) após as Avaliações de Conhecimento Posterior e de Percepção da Experiência (após as Etapas 4 e 5). A duração das pausas durante e imediatamente após o curso varia conforme decisão do controle do tempo livre, detalhado nas próximas seções.

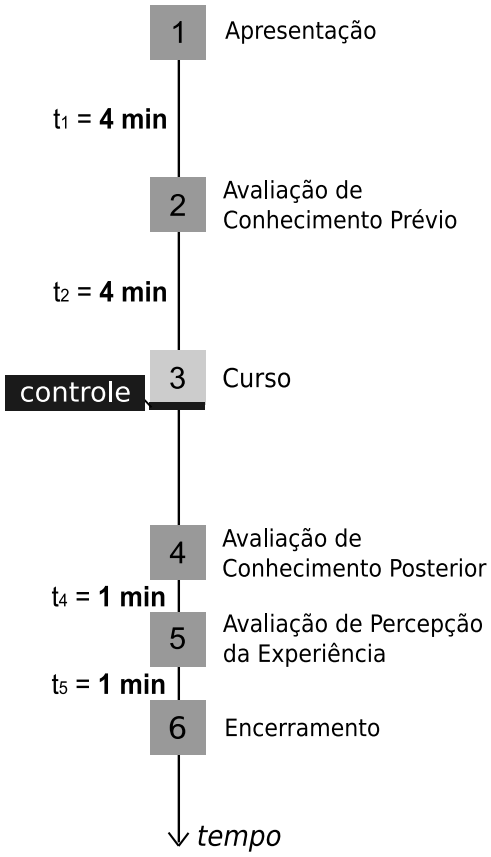

Figura 4.2: *Duração de tempo livre.*

4.1.2 Estados, ações e recompensas

Os **estados** constituem a base para as escolhas do agente [Sutton e Barto 1998]. No experimento, os estados são caracterizados pelos **módulos** do curso. A sequência de estados $s_1, s_2, s_3, ..., s_{15}$, referente a cada módulo, é fixa para o ambiente dinâmico.

Quanto às **ações**, o agente pode decidir por diminuir, manter ou aumentar a duração do tempo livre. No casos de diminuição e aumento, a pausa varia em 25% (a menos ou a mais)[2] do valor anterior. Portanto, numericamente, o conjunto de ações é: $A = \{0,75; 1; 1,25\}$. O controle inteligente escolhe

[2]Acredita-se que o valor escolhido (25%) causa variação perceptível e suave no tempo livre.

a ação a_i para determinação do tempo livre, conforme a Equação 4-1, onde t_i é o tempo livre atual, t_{i-1} é o tempo livre anterior. Baseado em estimativas, o tempo livre atual t_i possui limites mínimo, igual a um minuto, e máximo, igual ao tempo livre médio disponível por módulo, tendo em vista o tempo livre total estimado (100 min) e o tempo gasto até o momento.

$$t_i = a_i.t_{i-1}, \text{com} \quad a_i \epsilon A \tag{4-1}$$

Para avaliar as ações (escolhas), o agente utiliza a função-recompensa, ou seja, a média ponderada das notas, $N1$ e $N2$, nos exercícios, como instância da Equação 3-4. Os pesos, denotados por K e L, foram definidos como constantes com os seguintes valores: $K = 1$ e $L = 2$ (ver a Equação 4-2).

$$r = \frac{N1 + 2.N2}{3} \tag{4-2}$$

Tanto $N1$ quanto $N2$ são calculadas a partir da mesma fórmula, apresentada na Equação 3-5 e repetida abaixo, sendo p_j, a perda por tentativa, p_R, a perda de referência, ou seja, o somatório de perdas no pior caso, e ta, a tentativa do acerto. A Equação 4-4 mostra o cálculo do valor da perda por tentativa (p_j), onde T é o total de alternativas (tentativas possíveis) e $H(x)$, a função que caracteriza o tipo de resposta, Correta (C), Semelhante à Correta (SC) e Incorreta (I). Quando a perda no exercício é mínima, a nota é máxima, e vice-versa.

$$Nota = 1 - \frac{\sum_{j=1}^{ta} p_j}{p_R} \tag{4-3}$$

$$p_j = H(x).\frac{1}{T - j + 1}, \quad sendo \quad que \quad H(x) = \begin{cases} 0 & \text{se} \quad x = C \\ 0,5 & \text{se} \quad x = SC \\ 2 & \text{se} \quad x = I \end{cases} \tag{4-4}$$

Os valores de $H(x)$ e p_R foram definidos empiricamente. Para quatro alternativas e três tipos de respostas, existem nove cenários distintos de tentativas até o acerto (ver Tabela 4.1). O pior caso, Cenário 9, definiu o valor da perda de referência em $1,792$. Baseado nos cenários, o gráfico da Figura 4.3 mostra a perda no exercício (segunda parte da Equação 4-3) para os valores de $H(x)$ iguais a 0 (zero), $0,5$ e 2 para as respostas Correta (C), Semelhante à Correta (SC) e Incorreta (I), respectivamente, e $p_R = 1,792$.

Tabela 4.1: *Cenários das tentativas até o acerto baseados nos tipos de resposta.*

	Tentativas			
Cenário	1^a	2^a	3^a	4^a
1	C			
2	SC	C		
3	I	C		
4	I	SC	C	
5	SC	I	C	
6	I	I	C	
7	I	I	SC	C
8	I	SC	I	C
9	SC	I	I	C

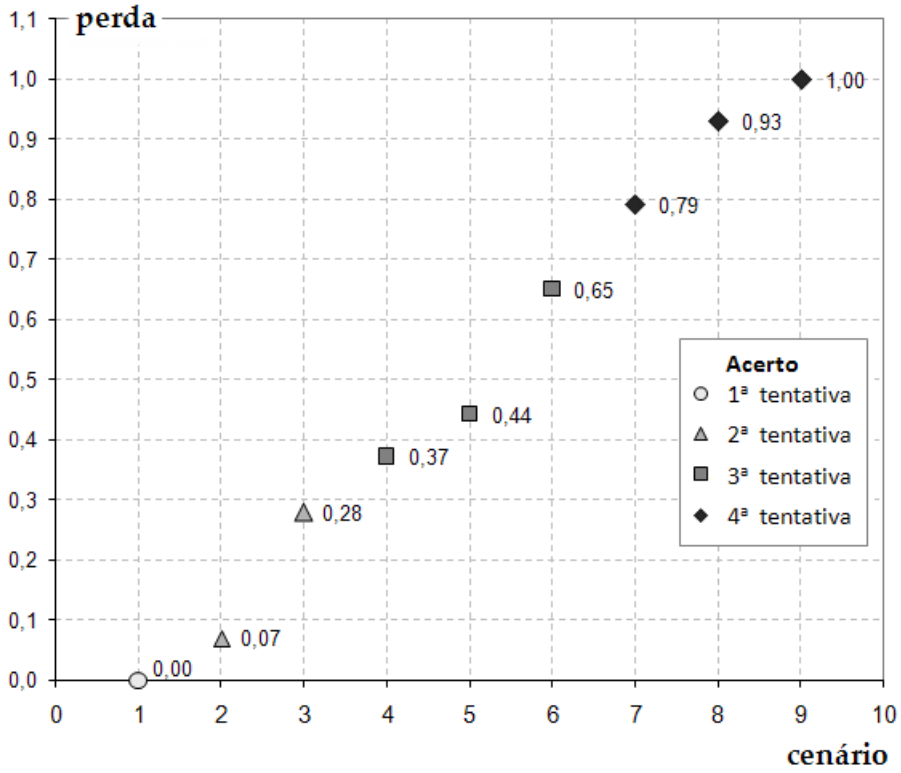

Figura 4.3: *Perda para os cenários de tentativas.*

4.1.3 Função-valor

A função-valor proposta é a média das recompensas observadas para cada ação. Para encorajar a exploração extensiva, o agente inicializa $Q(s, a)$ com um valor intermediário de r. Baseado nas recompensas mínima e máxima iguais a 0 (zero) e 1, respectivamente, $Q_0(s, a) = 0,5$ (conforme Equação 4-5) é otimizado, tendo em vista as simulações realizadas.

$$Q_0(s, a) = \frac{r_{min} + r_{max}}{2} \tag{4-5}$$

Ao final de cada passo do episódio (depois do exercício de revisão), o agente atualiza o valor da ação $Q(s, a)$ aplicada como um incremento da estimativa anterior, conforme mostra a instrução abaixo. A variável k é o número de vezes que a ação a foi selecionada, definindo assim o tamanho do passo.

$$Q(s, a) \leftarrow Q(s, a) + \frac{1}{k}[r - Q(s, a)] \tag{4-6}$$

4.1.4 Ajuste da Temperatura

O parâmetro τ, temperatura, do método Softmax contribui para o balanceamento entre exploração extensiva e intensiva, com decremento gradativo do seu valor. As primeiras simulações para ajuste da temperatura avaliaram a política para casos em que o aluno acertava sempre ou errava sempre. Nessas circuntâncias, o algoritmo comportou-se como um sistema aleatório (conforme mostra a Figura 4.4), sem convergir, pois errar ou acertar sempre indica a ausência de uma ação melhor, ou seja, que promova o acerto.

No gráfico da Figura 4.4, o eixo da abscissa representa os estados, enquanto o eixo da coordenada contém os valores do desvio padrão médio da política π. Para evidenciar a diferenciação das ações ao longo dos passos (transição de estados), o desvio padrão deve aumentar. Embora o comportamento observado fosse contrário ao desejado, o gráfico apontou o ponto em que, provavelmente, todas as ações já foram escolhidas, o Estado 8, s_8.

Para ajustar a temperatura e avaliar o comportamento do agente via simulações, fez-se necessário definir perfis distintos de aluno. Nesta pesquisa, a modelagem do problema da Aprendizagem por Reforço propõe três tipos de aprendiz: teórico, equilibrado e pragmático. O tipo teórico acerta quando o tempo livre diminui, logo na primeira tentativa, se o tempo livre mantém-se igual ao anterior, o aluno acerta na segunda tentativa, e erra consideravelmente (pior caso) se o tempo livre aumentar. Para o tipo pragmático, a situa-

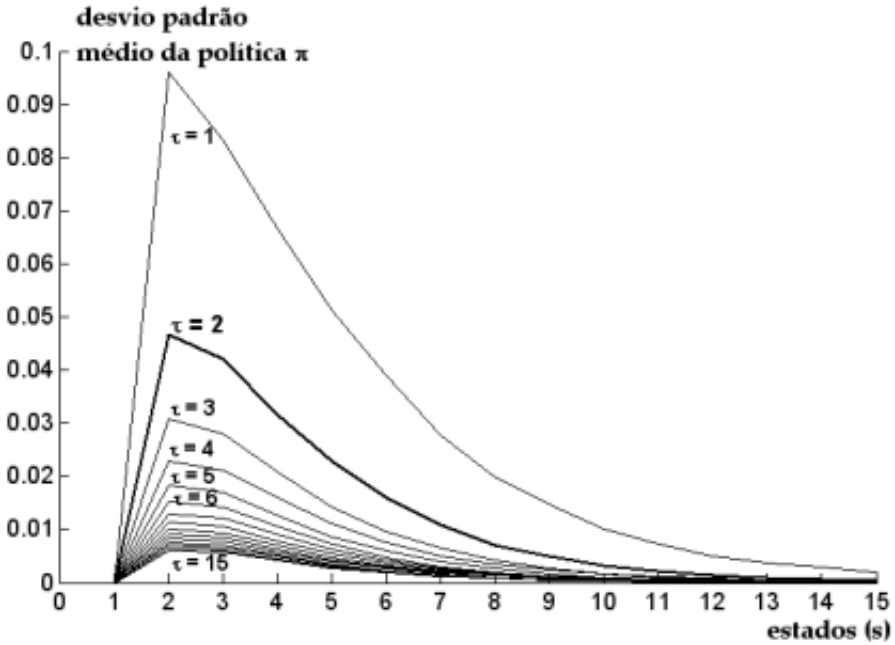

Figura 4.4: *Simulações (com 10000 iterações), em que aluno acerta sempre, para ajuste de temperatura τ variando de 1 a 15.*

ção é inversa. O tipo equilibrado acerta na primeira tentativa quando o tempo livre mantém-se e erra sempre que o tempo livre varia, seja para mais ou para menos, como mostra a Tabela 4.2.

Tabela 4.2: *Modelagem do problema da Aprendizagem por Reforço.*

Ação \ Perfil de aluno	Teórico	Equilibrado	Pragmático
Diminuir Tempo Livre	C	SC-I-I-C	SC-I-I-C
Manter Tempo Livre	SC-C ou I-C	C	SC-C ou I-C
Aumentar Tempo Livre	SC-I-I-C	SC-I-I-C	C

Tipos de resposta: C-Correta, SC-Semelhante à Correta e I-Incorreta

Conforme a modelagem proposta, as simulações evidenciaram o comportamento desejado para aprendizagem computacional, sendo a curva igual para os perfis "Teórico" e "Pragmático", como mostra o gráfico da Figura 4.5. Para temperatura inicial τ_0 igual a 2, as ações são praticamente equiprováveis. A temperatura τ com valores menores que $0,2$ apresentou diferenciação satisfatória entre as ações. Portanto a temperatura varia de 2 a $0,2$ do Estado 1 ao 8 e de $0,2$ a $0,1$ para os estados restantes (ver Figura 4.6). As razões do caimento nesses dois contextos, 1) Exploração Extensiva e 2) Exploração Intensiva, são aproximadamente $q_1 = -0,257$ e $q_2 = -0,014$.

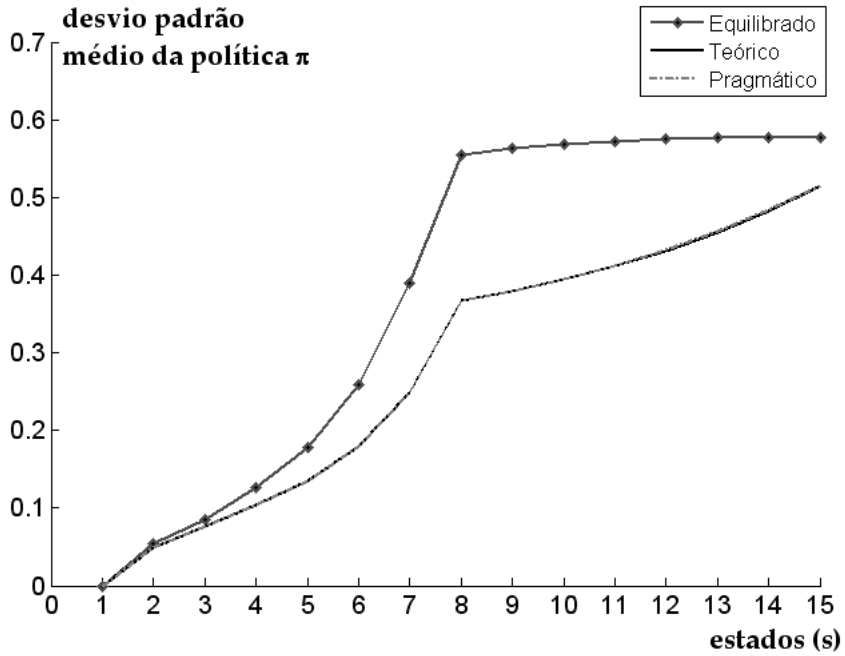

Figura 4.5: *Simulações (com 10000 iterações) para os perfis de aluno.*

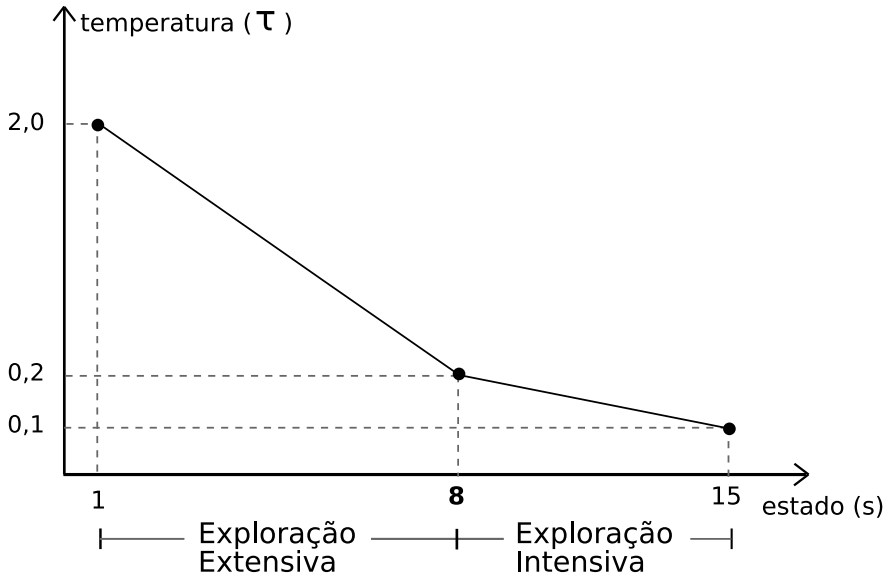

Figura 4.6: *Ajuste implementado da temperatura.*

4.1.5 Ambiente Tecnológico

O sistema tutor inteligente foi implementado em Java, linguagem de programação robusta e multiplataforma. Para rodar os vídeos durante a tutoria, o sistema utilizou a estrutura para mídia do Java, o *Java Media Framework* (JMF). Os vídeos foram codificados em "MJPEG", com áudio em "WAV", formatos (*codecs*) suportados pelo JMF no ambiente Linux.

Por ser um *software desktop* em Java, o sistema tutor pode ser executado em sistemas operacionais diferentes. Porém a coleta de dados ocorreu somente no Linux, na distribuição Ubuntu (versão 8.10), em concordância com o conteúdo do curso, "Introdução à Microinformática no Linux". Durante a coleta, o acesso à Internet foi bloqueado para evitar influências externas.

Todos os jogos padrões do Ubuntu, como "Xadrez", "Gnometris", "Tetravex", serviram como opções de entretenimento. No caso das músicas, escolhemos apenas músicas livres, disponíveis gratuitamente pela Internet. Na produção das vídeo-aulas, houve o mesmo rigor quanto aos direitos autorais, todas as ilustrações são de autoria do *designer* colaborador[3].

4.1.6 Amostra

A amostra contém 64 adolescentes do Estado de Goiás entre 14 e 17 anos. Todos os participantes frequentam escolas públicas (estaduais e federal) e possuem pouco ou nenhum conhecimento sobre Microinformática. Para possibilitar estudo comparativo, a amostra foi dividida em dois grupos de alunos com as mesmas características. A tabela abaixo apresenta os grupos, A e B, por cidade[4].

Tabela 4.3: *Grupos da amostra por cidade.*

Cidade \ Amostra	Grupo A	Grupo B
Inhumas	21	20
Santa Rosa de Goiás	10	10
Goiânia	1	1
Caturaí	0	1
Total	32	32

[3]Estudante de Artes Visuais, na Universidade Federal de Goiás. Foi selecionado para estagiar como *designer* gráfico nesta pesquisa em julho de 2008.

[4]Caturaí é uma cidade vizinha à Inhumas. O único participante de Caturaí estuda em Inhumas, onde foi selecionado.

4.2 Resultados

A hipótese básica deste trabalho apresenta uma solução provisória para o problema do controle de tempo livre em Sistemas Tutores Inteligentes. Para verificar a hipótese básica de que o controle via Aprendizagem por Reforço aumenta a retenção de conhecimento, utilizou-se o sistema tutor sem inteligência computacional para o grupo controle.

O controle de tempo livre aplicado diferencia os grupos da amostra. No grupo controle (Grupo A), o aluno escolhe manter ou variar a duração das pausas, **controle livre**. Enquanto no grupo experimental (Grupo B), o agente da Aprendizagem por Reforço decide por uma ação, **controle inteligente**. A partir da coleta de dados, as seções a seguir apresentam os resultados desta pesquisa, conforme análise (descritiva e inferencial) e interpretação dos dados.

4.2.1 Análise Descritiva

De forma descritiva, os dados são analisados em termos de mínimo, máximo, média e desvio padrão. Nas tabelas e gráficos, os termos "livre" e "inteligente" referem-se ao tipo de controle das pausas e, por consequência, ao grupo da amostra, A e B, respectivamente. Os dados são medidas do desempenho, do tempo gasto e da experiência do aluno nos sistemas tutores (livre e inteligente).

A apresentação dos dados evidenciam comportamento semelhante nos dois grupos. Os valores de desvio padrão, relativamente altos, apontam para heterogeneidade na amostra. Porém o rigor na seleção da amostra poderia inviabilizar a coleta de dados, devido às dificuldades enfrentadas em pesquisa com humanos, como alto índice de ausência na atividade, alocação de ambiente adequado, entre outras.

Quanto às variáveis observadas, vale ressaltar que as notas podem variar de 0 a 10. O ganho normalizado corresponde ao ganho obtido comparado ao ganho possível (diferença entre a nota máxima alcançável e a nota inicial), conforme mostra a Equação 4-7.

$$GanhoNormalizado = \frac{NotaFinal - NotaInicial}{NotaMaxima - NotaInicial} \tag{4-7}$$

Tanto no grupo controle, quanto no experimental, houve retenção de conhecimento (a nota final foi maior que a inicial). As Tabelas 4.4, 4.5 e 4.6 apresentam valores próximos para as notas inicial e final e para o ganho nos

dois grupos, inclusive em relação aos desvios padrões - aspecto importante para provar equivalência entre os grupos (amostras).

Tabela 4.4: *Estatística descritiva da* ***nota inicial****.*

Estatística\Amostra	Livre	Inteligente
Mínimo	0,67	1,33
Máximo	6,67	6,67
Média	3,88	4,13
Desvio Padrão	1,52	1,55

Tabela 4.5: *Estatística descritiva da* ***nota final****.*

Estatística\Amostra	Livre	Inteligente
Mínimo	2,00	2,67
Máximo	9,33	10,00
Média	6,56	6,71
Desvio Padrão	1,75	1,95

Tabela 4.6: *Estatística descritiva do* ***ganho normalizado****.*

Estatística\Amostra	Livre	Inteligente
Mínimo	0,0%	-0,1%
Máximo	90,0%	100,0%
Média	45,2%	45,2%
Desvio Padrão	23,4%	29,4%

No sistema tutor inteligente, o tempo total da atividade (desde a apresentação até o encerramento) foi ligeiramente menor, com os extremos, mínimo e máximo, mais próximos. Pode-se perceber maior homogeneidade no controle inteligente, tendo em vista desvio padrão menor. Em média, a atividade durou duas horas e 31 minutos no controle inteligente e duas horas e 34 minutos no controle livre (ver Tabela 4.7).

Os sistemas tutores avaliam a percepção da experiência na atividade mediante a aplicação de um questionário com seis questões. As respostas da primeira pergunta (“Em relação ao curso, que etapa você considera mais interessante?”) correspondem às etapas dos módulos: vídeo-aula, exercício, sugestão prática, tempo livre e exercício de revisão. A Figura 4.7 mostra os percentuais de escolha das etapas.

Tabela 4.7: *Estatística descritiva do* ***tempo total*** *da atividade.*

Estatística\Amostra	Livre	Inteligente
Mínimo	1º52'1"	1º57'55"
Máximo	3º26'6"	3º22'20"
Média	2º33'55"	2º31'29"
Desvio Padrão	28'24"	25'29"

Nos dois grupos, a maioria dos alunos considerou a vídeo-aula como a etapa mais interessante. No grupo experimental, as outras fases apresentaram distribuição uniforme, diferentemente do grupo controle. Tanto o tempo livre quanto as etapas próximas a ele, sugestão prática e exercício de revisão, destacaram-se no controle inteligente, indicando melhor percepção da pausa como componente da estratégia de ensino.

A sugestão prática, planejada para promover postura ativa do aluno, obteve cinco vezes mais escolhas no controle inteligente, sendo a segunda opção apontada pelos alunos. Com o triplo da frequência das respostas no grupo experimental (comparado ao grupo controle), o tempo livre foi considerado parte integrante do curso, e o exercício de revisão, apresentado após o tempo livre, recebeu o mesmo percentual de escolhas da sugestão prática, 15,6%, no controle inteligente.

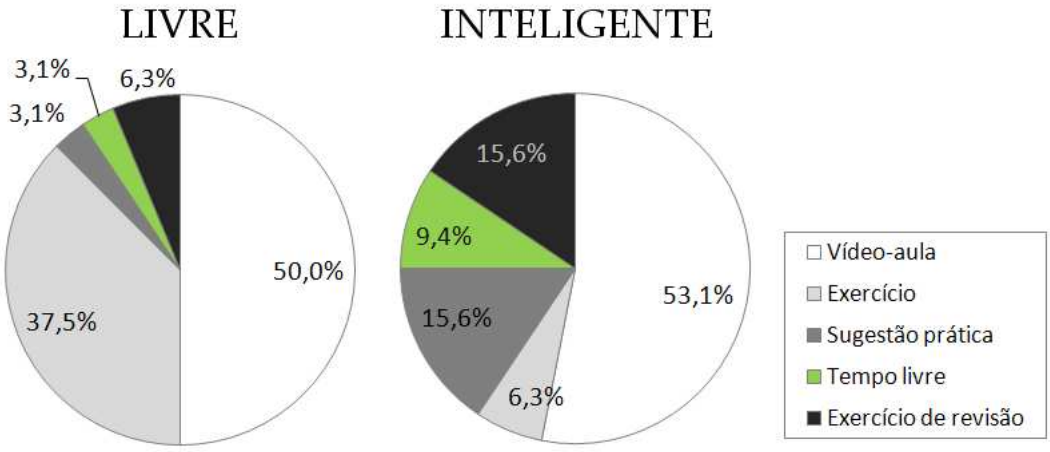

Figura 4.7: *Etapa do curso mais interessante.*

Sobre a duração do tempo livre, houve maior satisfação no grupo controle, pois o próprio aluno decidia manter ou variar o tamanho da pausa. De forma geral, os dois grupos consideraram o tempo suficiente, com mais de 60% das respostas (ver Figura 4.8).

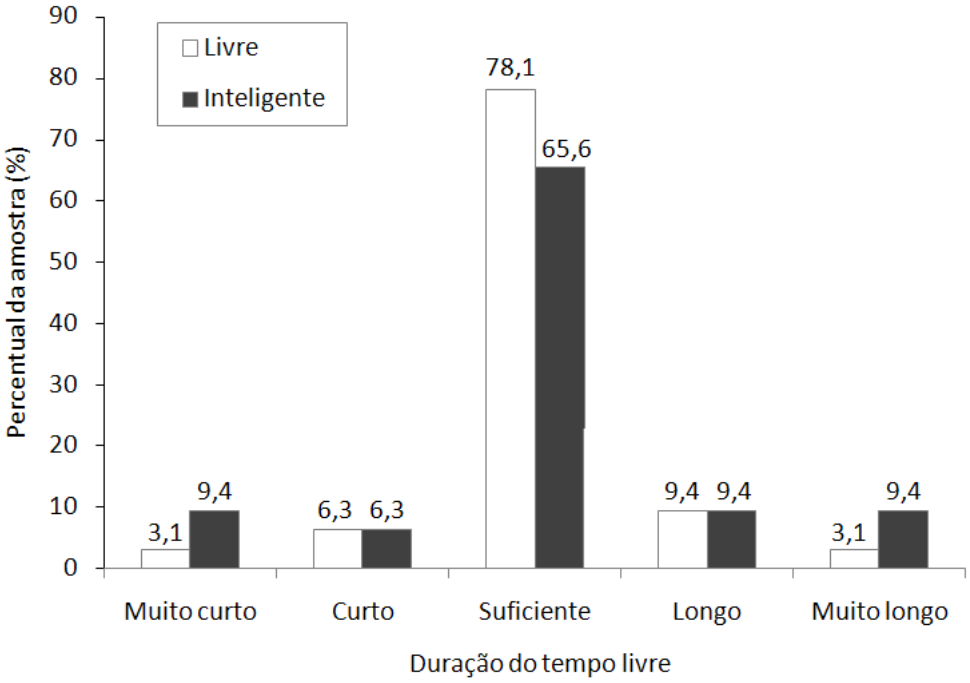

Figura 4.8: *Avaliação da duração do tempo livre.*

O aluno indicou o quanto aprendeu do curso, revelando sua percepção da aprendizagem. No controle inteligente, houve casos isolados de resposta "pouco" e "nada do conteúdo" (ver Tabela 4.8). Nas outras opções, os grupos apresentaram opinião semelhante, com pequena diferença no item "todo o conteúdo", favorável ao controle inteligente, como mostra a Figura 4.9.

Tabela 4.8: *Frequência de respostas sobre a percepção da aprendizagem.*

Resposta \ Amostra	Livre	Inteligente
Nada do conteúdo	0	1
Pouco	0	1
Metade do conteúdo	5	5
Muito	21	18
Todo o conteúdo	6	7

Para a pergunta "O que você mais gostou de fazer no tempo livre?", as opções de entretenimento ("Jogar" e "Ouvir músicas") foram apontadas por mais de 70% dos alunos em ambos os grupos. No controle inteligente, a alternativa "Fazer a sugestão prática do curso" obteve percentual um pouco maior que no controle livre, conforme mostra a Figura 4.10.

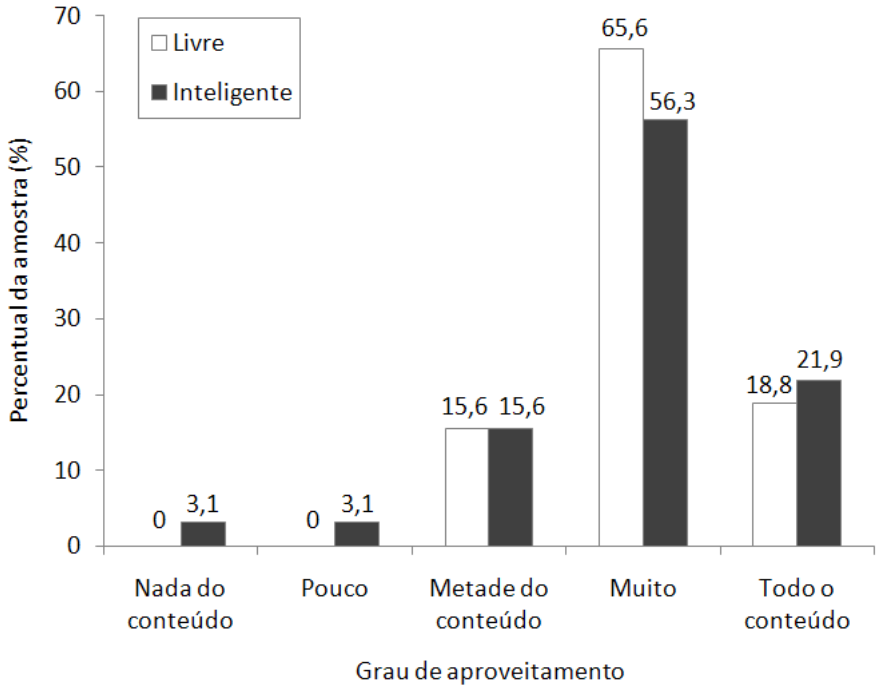

Figura 4.9: *Percepção da aprendizagem.*

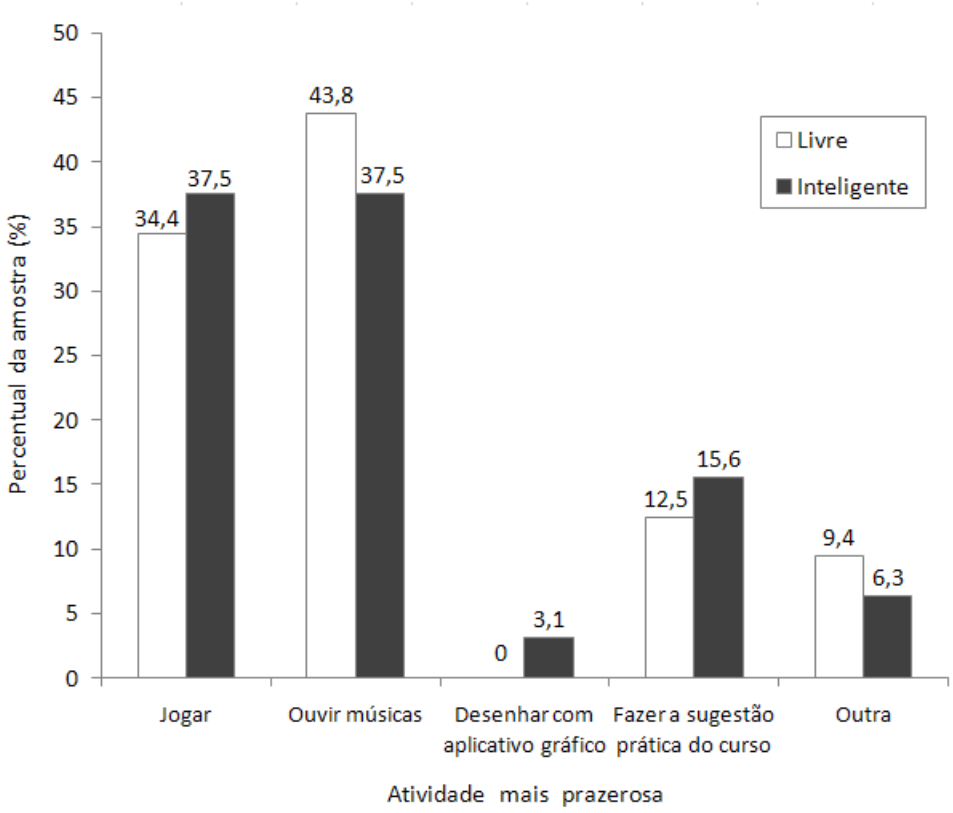

Figura 4.10: *Preferência de uso do tempo livre.*

O cenário sobre a percepção de toda a atividade, mostrado na Figura 4.11, revela maior agradabilidade no sistema tutor livre. Este gráfico condiz com o apresentado na Figura 4.8, sendo esta situação justificada pelo tipo de controle, livre, onde o aluno decide se e como variar a duração da pausa.

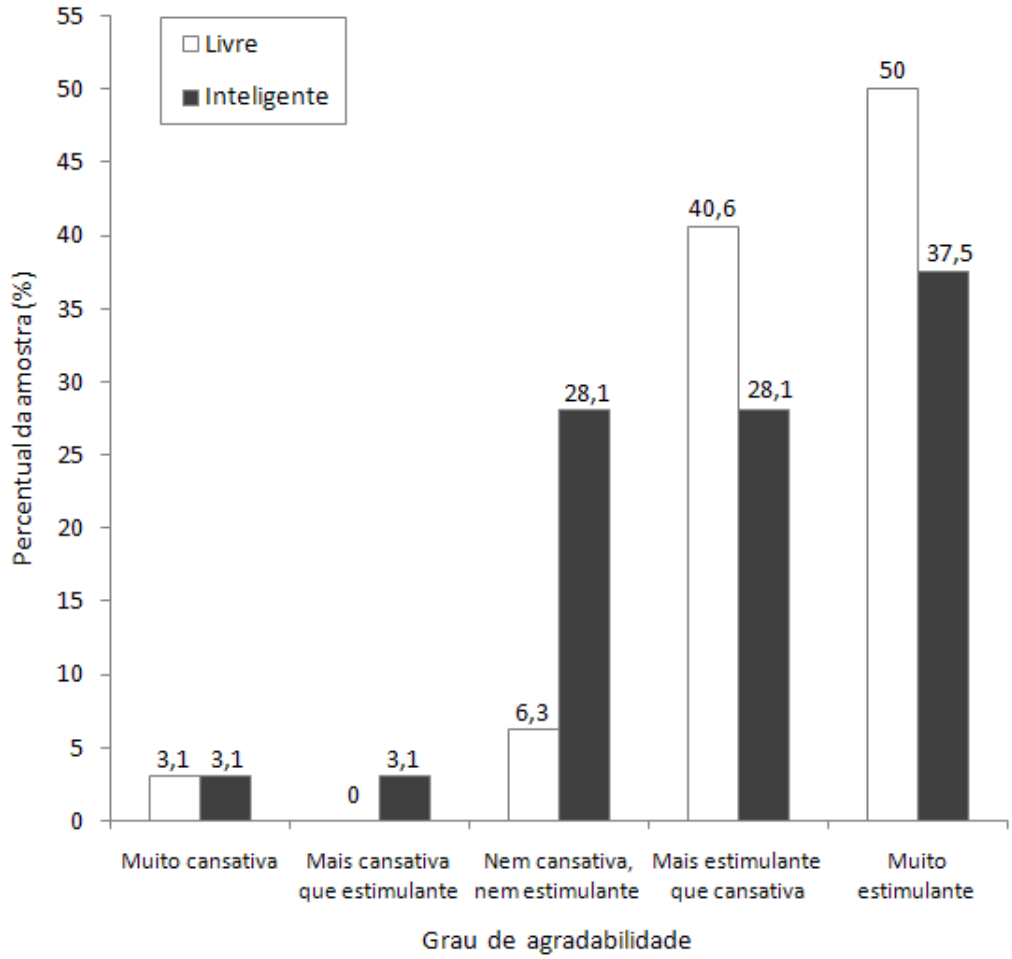

Figura 4.11: *Percepção da atividade inteira.*

De forma unânime, todos gostaram do sistema tutor. O aluno deveria indicar o quanto gostou da atividade escolhendo uma das alternativas: a) Gostei demais, b) Gostei, c) Nem gostei, nem não gostei, d) Não gostei, e) Não gostei de nada, odiei. Somente as opções “a” e “b” foram selecionadas (ver Figura 4.12), com predomínio para a resposta “Gostei demais” nos dois grupos.

No encerramento, o aluno respondia se desejava continuar ou não no computador e digitava o motivo. Nos dois grupos, os alunos preferiram sair da sala, afirmando a necessidade de ir embora (na maioria dos casos). O controle inteligente apresentou maior percentual na opção de continuar no computador, conforme Figura 4.13.

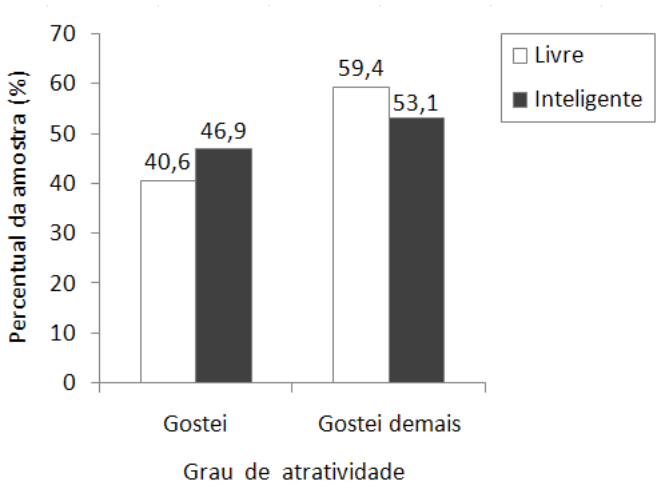

Figura 4.12: *Gosto pela atividade.*

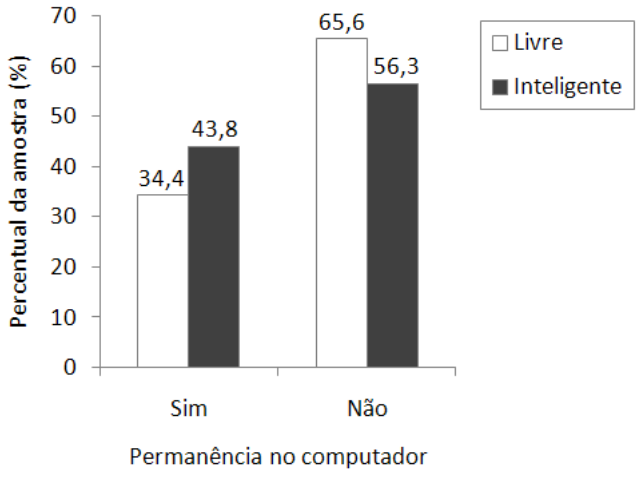

Figura 4.13: *Permanência no computador.*

4.2.2 Análise Inferencial

As relações estatísticas (média, desvio padrão e frequência) da análise descritiva representam as características (desempenho, tempo total de atividade) da amostra e fundamentam as conclusões. A partir de testes estatísticos, como Teste t (*Student*) e Qui-Quadrado, a análise inferencial generaliza as conclusões para população [Costa 1998]. Esta seção apresenta provas de hipóteses para verificar se há diferenças significativas entre as variáveis observadas nos grupos controle e experimental.

Nas tabelas a seguir, H_a é a hipótese alternativa, ou seja, a conjectura que poderá substituir a hipótese nula (H_0), caso H_0 seja rejeitada (quando existem diferenças significativas). Neste trabalho, a prova da hipótese nula possui nível de significância, denotado por α, igual a 5%. Nos testes estatísticos, α efetivo é o nível de significância observado e corresponde ao valor-p (do Inglês, *p-value*). Para que H_0 seja rejeitada, o valor do α efetivo deve ser menor ou igual a 5%, α de referência, .

O teste t (*Student*) pareado comparou a nota inicial com a nota final dos participantes, a fim de provar o aumento significativo no desempenho do aluno (ver gráfico da Figura 4.14). A Tabela 4.9 mostra a estatística inferencial de toda amostra, enquanto as Tabelas 4.10 e 4.11 apresentam a estatística por grupo, livre e inteligente, respectivamente. Nas três situações, a hipótese nula foi rejeitada. Os sistemas tutores colaboraram na aprendizagem dos alunos, pois houve retenção de conhecimento com a tutoria.

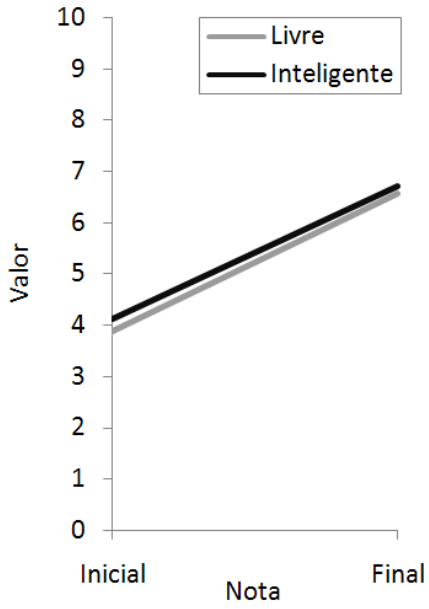

Figura 4.14: *Desempenho dos alunos na atividade.*

Tabela 4.9: *Estatística inferencial do* ***desempenho geral*** *- Teste t pareado.*

$H_a : \mu_{notaFinal} > \mu_{notaInicial}$

Estatística \ Nota	Inicial	Final
Média	4,00	6,64
Desvio Padrão	1,53	1,84
Observações	64	64
t observado	13,29	
α efetivo unicaudal	0,00	
t crítico	1,67	

Tabela 4.10: *Estatística inferencial do* ***desempenho da amostra "livre"*** *- Teste t pareado.*

$H_a : \mu_{notaFinal} > \mu_{notaInicial}$

Estatística \ Nota	Inicial	Final
Média	3,88	6,56
Desvio Padrão	1,52	1,75
Observações	32	32
t observado	10,47	
α efetivo unicaudal	0,00	
t crítico	1,7	

Os grupos submetidos aos tipos de controle de tempo livre são amostras independentes e equivalentes. Portanto a média das notas iniciais na população é aproximadamente igual tanto no controle inteligente quanto no livre, visto que H_0 não foi rejeitada, conforme mostra a Tabela 4.12.

Ao contrário, esperava-se que a nota final e o ganho do sistema proposto fossem maiores (H_a), porém os valores observados (ver Tabelas 4.13 e 4.14) provam que H_0 é verdadeira, ou seja, não houve diferenças significativas entre os grupos após a tutoria.

Tabela 4.11: *Estatística inferencial do* ***desempenho da amostra "inteligente"*** *- Teste t pareado.*

$H_a : \mu_{notaFinal} > \mu_{notaInicial}$		
Estatística \ Nota	Inicial	Final
Média	4,13	6,71
Desvio Padrão	1,55	1,95
Observações	32	32
t observado	8,43	
α efetivo unicaudal	0,00	
t crítico	1,7	

Tabela 4.12: *Estatística inferencial da* ***nota inicial*** *- Teste t com amostras independentes.*

$H_a : \mu_{notaInicialInteligente} \neq \mu_{notaInicialLivre}$		
Estatística \ Amostra	Livre	Inteligente
Média	3,88	4,13
Desvio Padrão	1,52	1,55
Observações	32	32
t observado	0,65	
α efetivo bicaudal	0,52	
t crítico	2	

Quanto ao tempo total de tutoria, a Tabela 4.15 apresenta a estatística inferencial baseada nas observações, onde H_0 não foi rejeitada. Em ambos os grupos, os alunos gastam, em média, o mesmo tempo para terminarem a tutoria - tempo gasto da Etapa 1, Apresentação, à Etapa 6, Encerramento.

Os valores observados sobre a percepção da experiência de tutoria foram insuficientes para a prova de Qui-Quadrado. Muitas células apresentaram frequências inferiores ao necessário. Somente a última questão obteve quantia suficiente de dados (ver Figura 4.12), onde o teste de Qui-Quadrado evidenciou α efetivo igual a $0,801$, H_0 não-rejeitada. No encerramento, a permanência do aluno no computador foi a mesma nos grupos, com α efetivo igual a $0,609$ (H_0 verdadeira).

Tabela 4.13: *Estatística inferencial da **nota final** - Teste t com amostras independentes.*

$H_a : \mu_{notaFinalInteligente} > \mu_{notaFinalLivre}$

Estatística \ Amostra	Livre	Inteligente
Média	6,56	6,71
Desvio Padrão	1,75	1,95
Observações	32	32
t observado	0,31	
α efetivo unicaudal	0,38	
t crítico	1,67	

Tabela 4.14: *Estatística inferencial do **ganho normalizado** - Teste t com amostras independentes.*

$H_a : \mu_{ganhoInteligente} > \mu_{ganhoLivre}$

Estatística \ Amostra	Livre	Inteligente
Média	45,2%	45,2%
Desvio Padrão	23,4%	29,4%
Observações	32	32
t observado	0,05	
α efetivo unicaudal	0,48	
t crítico	1,67	

Tabela 4.15: *Estatística inferencial do **tempo total** de atividade - Teste t com amostras independentes.*

$H_a : \mu_{tempoTotalInteligente} < \mu_{tempoTotalLivre}$

Estatística \ Amostra	Livre	Inteligente
Média	9235.31’	9088.97’
Desvio Padrão	1704,23’	1529,1’
Observações	32	32
t observado	-0,36	
α efetivo unicaudal	0,36	
t crítico	-1,67	

4.2.3 Discussão

Conforme as condições experimentais, a hipótese básica desta pesquisa não foi corroborada, ou seja, a Aprendizagem por Reforço para controlar o tempo livre manteve os mesmos níveis de retenção de conhecimento que o controle humano. As hipóteses secundárias são desconsideradas, pois dependem do fortalecimento da hipótese básica para serem analisadas. De forma geral, não houve diferenças significativas entre os tipos de controle do tempo livre, inteligente (com Aprendizagem por Reforço) e livre (decisão do aluno).

No grupo experimental ("inteligente"), os estudantes perceberam melhor o tempo livre como componente da estratégia de ensino, conforme observações feitas durante a coleta dos dados e confirmadas pela análise descritiva (ver Figura 4.7). O sistema tutor livre permitiu ao aluno controlar a duração da pausa, desviando, contudo, parte de sua atenção para o processo de escolha. Assim, o grupo controle considerou o tempo livre e as etapas imediatamente antes (sugestão prática) e depois (exercício de revisão) alheios ao curso.

Para o grupo controle ("livre"), a condição de liberdade para alterar ou manter a duração da pausa provocou maior satisfação na atividade, verificada pela alta frequência de respostas positivas (tempo livre suficiente, atividade estimulante). Ao controlar o tamanho da pausa, o aprendiz percebe-se como responsável por esta etapa do sistema tutor.

Referente à prática no tempo livre, os estudantes exploravam as funcionalidades computacionais mediante acompanhamento de um tutor humano. Quando se envolveram em atividades práticas, os alunos apresentaram aumento no interesse pelo curso, principalmente, durante o tempo livre, na participação ativa no processo da aprendizagem.

Na transmissão de conhecimento, o uso de vídeo-aula contribuiu para desempenho semelhante nos grupos controle e experimental, além de ser apontada como etapa mais interessante pela maioria dos alunos. No trabalho de [Martins et al. 2007], a Aprendizagem por Reforço foi empregada na seleção de conteúdos "estáticos" para os aprendizes no cenário tradicional de uma única sessão, obtendo ganho normalizado médio de 52,61%. O presente trabalho aliou o uso de pequenas vídeo-aulas e a inserção de tempo livre, atingindo ganhos equivalentes não apenas para o sistema inteligente, mas inclusive para o tutor livre.

Quanto à técnica de Aprendizagem por Reforço, a instabilidade do ambiente penalizou a aprendizagem computacional, visto que muitos fatores

afetam a interação com o aluno, inclusive emocionais[5]. Além disso, mesmo em situações estáveis, o agente inteligente necessita de um número maior de passos no episódio para obter melhor performance. Apenas em quatro casos, o sistema proposto realmente convergiu, com desvio padrão da política no último estado maior que $0,4$, conforme dados apresentados no Apêndice A.3, Tabelas A.1, A.2, A.3 e A.4.

Os gráficos abaixo mostram a evolução da política durante a tutoria, a partir da transição de estados. Os dados do aluno A1 (ver Figura 4.15) apresentam diferenciação mais significativa entre as ações (desvio padrão no último estado igual a $0,484$). Com desvio padrão menor, $0,237$, o gráfico da Figura 4.16 sugere convergência da Aprendizagem por Reforço, no entanto as escolhas do agente alternaram-se entre diminuir e aumentar a duração da pausa. Tal situação justifica-se pela probabilidade considerável, $19,64\%$, para a ação "Aumentar TL", onde TL é a sigla para Tempo Livre. Conforme Figura 4.17, não houve convergência para o aluno A32 (desvio padrão igual a $0,039$).

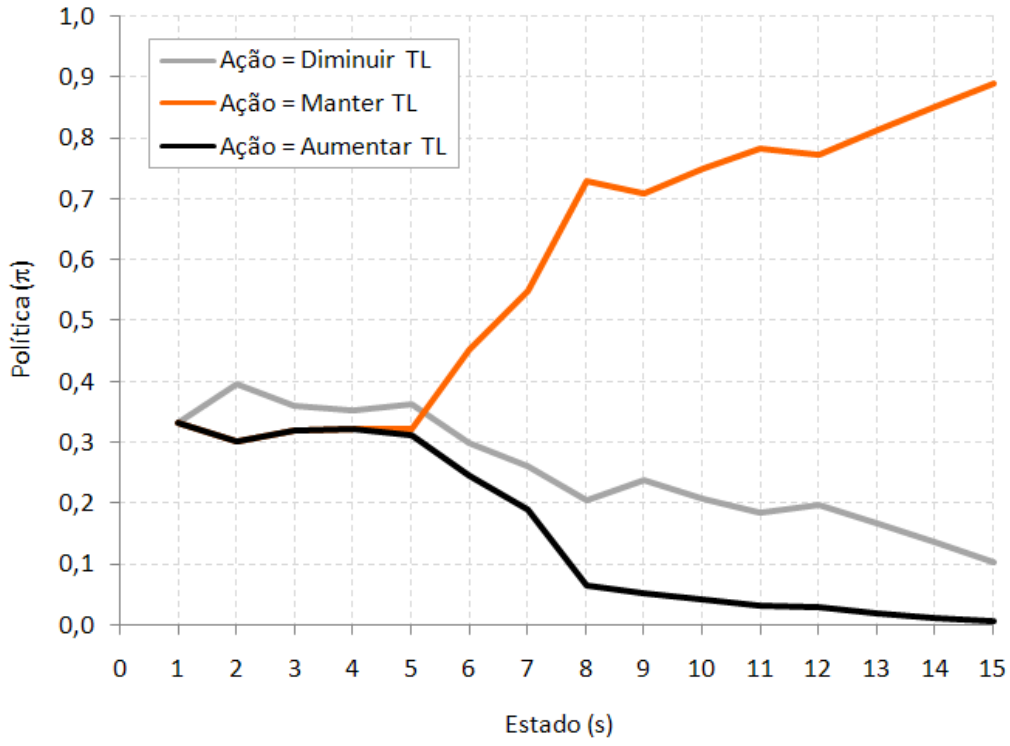

Figura 4.15: *Política do aluno A1.*

[5]Durante toda tutoria, o aprendiz tinha ciência do horário e do seu percurso no curso. Portanto, em alguns casos, o controle inteligente estava convergindo para aumentar a pausa e o aluno mostrava-se preocupado com o tempo longo (devido a outros compromissos).

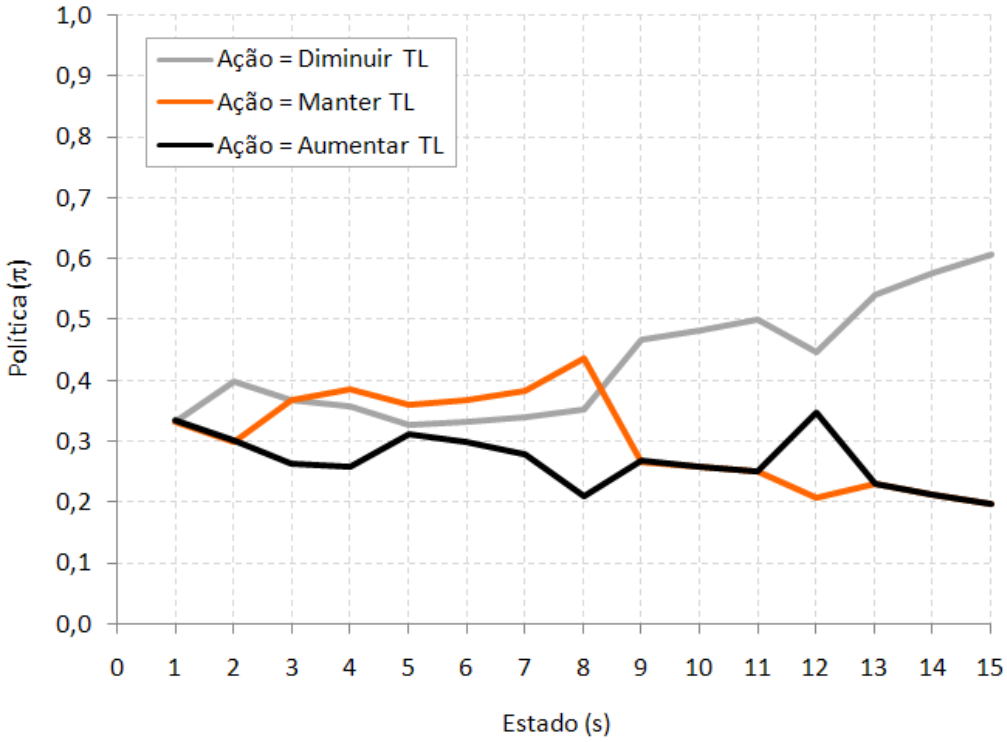

Figura 4.16: *Política do aluno A17.*

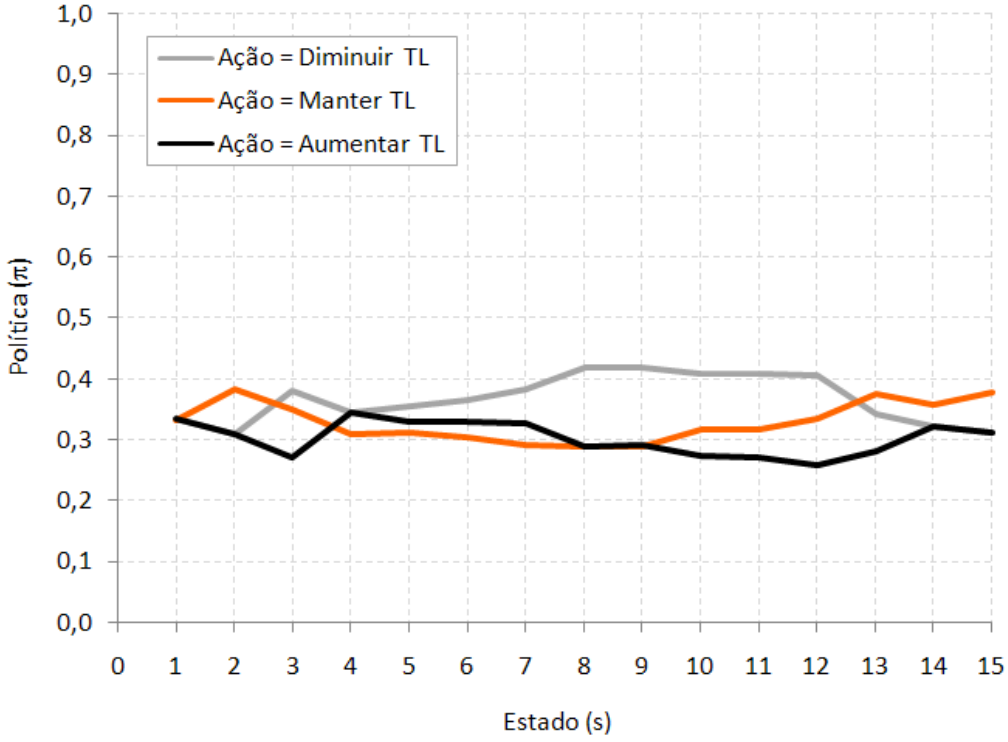

Figura 4.17: *Política do aluno A32.*

CAPÍTULO 5

Conclusão

5.1 Considerações gerais

Esta pesquisa introduz a inserção de pausas em Sistemas Tutores Inteligentes, ressaltando seu forte potencial como componente na estratégia de ensino, aliada ao uso de pequenas vídeo-aulas. As condições do delineamento foram insuficientes para comprovação da hipótese experimental, ou seja, a possível influência do tipo de controle do tempo livre sobre o desempenho do aluno.

A participação voluntária dos adolescentes na coleta de dados foi um ponto difícil da pesquisa, devido à alta taxa de ausência na atividade e a problemas disciplinares. No entanto, os estudantes nessa faixa etária (de 14 a 17 anos) naturalmente interagiram com o computador, mesmo quando era o primeiro contato. Os alunos adquiriram autonomia no uso do computador facilmente, contando com a ajuda do monitor apenas no início.

5.2 Principais contribuições

Cientificamente, este trabalho de prospecção apontou várias situações, propostas científicas, baseadas na inserção de tempo livre em STIs. A partir desta pesquisa, muitas outras poderão surgir como resposta às questões em aberto, tais como: o impacto do tempo livre para minimizar a tarefa como um todo, o uso de atividades práticas no tempo livre para complementar a tutoria, a influência do tempo livre na internalização do conteúdo, a maximização do tempo livre para alunos interessados em jogos interativos, entre outros.

A técnica de Inteligência Computacional, Aprendizagem por Reforço, produziu resultados semelhantes aos do controle humano do tempo livre, evidenciando, portanto, a capacidade de agentes inteligentes tomarem decisões

coerentes em sistemas tutores. O processo de convergência ocorreu mediante a preferência por uma ação (com base no desempenho), em conformidade com o conhecimento teórico da técnica utilizada.

Quanto ao experimento, os sistemas tutores contribuíram na capacitação de adolescentes no uso do computador, ao oportunizar a interação com o *software* tutor e com o ambiente Linux. Nesse contexto, o cenário proposto permite investigações sob abordagens distintas, como a Análise Experimental do Comportamento (tempo livre como elemento reforçador), o Construtivismo (postura ativa do aprendiz na construção do conhecimento durante o tempo livre), o Cognitivismo (influência do tempo livre na internalização do conteúdo) e outras.

5.3 Trabalhos futuros

A conjectura testada neste trabalho refere-se ao controle do tempo livre, com e sem inteligência computacional. Em trabalhos futuros, sugere-se a investigação sobre a eficácia do tempo livre na retenção de conhecimento, ou seja, estudo comparativo de sistemas tutores caracterizados pela presença ou ausência da pausa.

Em relação ao experimento, propõe-se que o sistema informe ao estudante a duração (quantidade de minutos) do tempo livre imediatamente antes de iniciá-lo, a fim de evitar possível ansiedade. Como opções para pausa no computador, atividades interativas relacionadas ao conteúdo, tais como palavras cruzadas, jogos educativos, aprimorariam a tutoria. Além disso, as sugestões práticas podem ser disponibizadas para consulta no formato impresso ou digital, num diretório na área de trabalho. Nos questionários para avaliar o conhecimento prévio e posterior, é recomendável colocar a opção “Não sei.” a fim de contribuir para veracidade dos resultados.

Quanto ao curso, um novo experimento poderia usar a sequência das etapas fundamentada na Aprendizagem Baseada em Problema (do Inglês, *Problem-based Learning*) [Jacinto e Oliveira 2007]. Em ordem, as etapas sugeridas são: 1) Problema, 2) Tempo Livre, 3) Exercício Procedural (ligado ao problema), 4) Vídeo-aula, 5) Exercício Conceitual. Assim cada módulo começaria com uma dúvida a ser respondida, o problema (Etapa 1). E os ensinamentos das vídeo-aulas provavelmente seriam mais proveitosos, pois o contato inicial (Etapas 1, 2 e 3) com o assunto prepararia o aluno para assimilação da teoria (Etapas 4 e 5).

A Aprendizagem por Reforço poderia ser modelada ainda com base em comportamento não-verbal, como ansiedade física e expressões faciais [Rodrigues e Carvalho 2005] [Sarrafzadeh et al. 2008]. Nesse sentido, o agente inteligente ofereceria tempo livre ao aprendiz apenas quando necessário (ao diagnosticar cansaço, por exemplo), ao invés de existirem pontos predeterminados para tal.

Referências Bibliográficas

(Araújo 2000) ARAÚJO, I. Z. de. *Controle de Sinal Fisiológico Humano Baseado em Aprendizado por Reforço: Uma Abordagem Competitiva*. Dissertação (Mestrado) — Universidade Federal de Goiás, Goiás, 2000.

(Barbosa 2008) BARBOSA, A. F. *Pesquisa sobre Uso das Tecnologias da Informação e Comunicação no Brasil: TIC Domicílios e TIC Empresas 2008*. São Paulo, 2008. Disponível em: <http://www.cetic.br/tic/2008/index.htm>.

(Baum 2005) BAUM, W. M. *Understanding behaviorism: Behavior, Culture and Evolution*. 2. ed. Malden, MA, USA and Oxford, OX, UK and Carlton, Victoria, Australia: Blackwell, 2005.

(Bernstein, Luthans e Welsh 1993) BERNSTEIN, D. J.; LUTHANS, F.; WELSH, D. H. B. Application of the premack principle of reinforcement to the quality performance of service employees. *Journal of Organizational Behavior Management*, 1993.

(Burns e Capps 1988) BURNS, H. L.; CAPPS, C. G. Foundations of intelligent tutoring systems. In: _____. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1988. cap. Foundations of Intelligent Tutoring Systems: An Introduction, p. 1–20.

(Candau 1969) CANDAU, V. M. *Ensino Programado: uma nova tecnologia didática*. Rio de Janeiro: Iter Edições, 1969.

(Catania 1999) CATANIA, A. C. *Aprendizagem: Comportamento, Linguagem e Cognição*. 4. ed. Porto Alegre: Artes Médicas Sul, 1999.

(Cohen 1995) COHEN, P. R. *Empirical Methods for Artificial Intelligence*. Cambridge, Massachusetts and London, England: The MIT Press, 1995.

(Costa 1998) COSTA, S. F. *Introdução Ilustrada à Estatística*. 3. ed. São Paulo: Editora HARBRA ltda., 1998.

(Eberspächer e Kaestner 2009) EBERSPÄCHER, H. F.; KAESTNER, C. A. A. *A Arquitetura de um Sistema de Autoria para Construção de Tutores Inteligentes Hipermídia e seu Posicionamento na Informática Educativa*. Paraná, 2009. Disponível em: <http://www.c5.cl/ieinvestiga/actas/ribie98/207.html>. Acesso em: 14 julho 2009.

(Folha 2008) FOLHA, Online. *Educação e conectividade fazem Brasil "patinar"em índice de convergência digital*. São Paulo, 2008. Disponível em: <http://www1.folha.uol.com.br/folha/informatica/ult124u467566.shtml>. Acesso em: 14 janeiro 2009.

(Foundation 2009) FOUNDATION, Free Software. *What is free software and why is it so important for society?* New York, 2009. Disponível em: <http://www.fsf.org/about/what-is-free-software>. Acesso em: 20 maio 2009.

(Frénay e Saerens 2009) FRÉNAY, B.; SAERENS, M. QL2, a simple reinforcement learning scheme for two-player zero-sum markov games. *Neurocomputing*, v. 72, n. 7–9, p. 1494–1507, March 2009.

(Geiger 1996) GEIGER, B. A time to learn, a time to play: Premack's principle applied in the classroom. *American Secondary Education*, v. 25, p. 2–6, 1996.

(Goulart 2007) GOULART, I. B. *Psicologia da Educação*: Fundamentos teóricos e aplicações à prática pedagógica. 13. ed. Petrópolis: Editora Vozes, 2007.

(Haykin 2001) HAYKIN, S. *Redes Neurais: Princípios e Prática*. 2. ed. Porto Alegre: Bookman, 2001.

(Holland e Skinner 1975) HOLLAND, J. G.; SKINNER, B. F. *A Análise do Comportamento*. São Paulo: Ed. da Universidade de São Paulo, 1975.

(Jacinto e Oliveira 2007) JACINTO, A. S.; OLIVEIRA, J. M. P. de. Uma investigação de STI que emprega a PBL de forma individual. *Anais do XVIII Simpósio Brasileiro de Informática na Educação*, p. 432–440, 2007.

(Kaelbling, Littman e Moore 1996) KAELBLING, L. P.; LITTMAN, M. L.; MOORE, A. W. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, v. 4, p. 237–285, 1996.

(Kalidindi e Bowman 2007) KALIDINDI, K.; BOWMAN, H. Using e-greedy reinforcement learning methods to further understand ventromedial prefrontal patients' deficits on the iowa gambling task. *Neural Networks*, v. 20, p. 676–689, 2007.

(Kasabov 1998) KASABOV, N. K. *Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering*. 2. ed. Cambridge, Massachusetts and London, England: Bradford Book/MIT Press, 1998.

(Lastres et al. 2002) LASTRES, H. M. M. et al. Desafios e oportunidades da era do conhecimento. *São Paulo em Perspectiva*, v. 16, n. 3, p. 60–66, 2002.

(Leung e Li 2007) LEUNG, E. W. C.; LI, Q. An experimental study of a personalized learning environment through open-source software tools. *IEEE Transactions on Education*, v. 50, n. 4, p. 331–337, 2007.

(Martins et al. 2007) MARTINS, W. et al. Tutoriais inteligentes baseados em aprendizado por reforço: Concepção, implementação e avaliação empírica. *Anais do XVIII Simpósio Brasileiro de Informática na Educação*, p. 617–626, 2007.

(Martins et al. 2004) MARTINS, W. et al. A novel hybrid intelligent tutoring system and its use of psychological profiles and learning styles. *Lecture Notes on Computer Science*, v. 3220, p. 830–832, 2004.

(Matignon e Fort-Piat 2007) MATIGNON, G. J. L. L.; FORT-PIAT, N. L. Hysteretic q-learning : an algorithm for decentralized reinforcement learning in cooperative multi-agent teams. *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems San Diego, CA, USA*, 2007.

(Mitchell e Stoffelmayr 1973) MITCHELL, W. S.; STOFFELMAYR, B. E. Application of the premack principle to the behavioral control of extremely inactive schizrophrenics. *Journal of Applied Behavior Analysis*, v. 6, n. 3, p. 419–423, 1973.

(Murray 1999) MURRAY, T. Authoring intelligent tutoring systems: An analysis of the state of the art. *International Journal of Artificial Intelligence in Education*, v. 10, p. 98–129, 1999.

(Neto 2002) NETO, M. B. de C. Análise do comportamento: behaviorismo radical, análise experimental do comportamento e análise aplicada do comportamento. *Interação em Psicologia*, v. 6, n. 1, p. 13–18, 2002.

(Osborne 1969) OSBORNE, J. G. Free-time as a reinforcer in the management of classroom behavior. *Journal of Applied Behavior Analysis*, v. 2, n. 2, p. 113–118, 1969.

(Portal 2009) PORTAL, Inclusão Digital. *Programas*. Brasília, 2009. Disponível em: <http://www.inclusaodigital.gov.br/inclusao/outros-programas>. Acesso em: 21 maio 2009.

(Premack 1959) PREMACK, D. Toward empirical behavior laws: I. positive reinforcement. *Psychological Review*, v. 66, p. 219–233, 1959.

(Rodrigues e Carvalho 2005) RODRIGUES, L. M. L.; CARVALHO, M. STI-I: Sistemas tutoriais inteligentes que integram cognição, emoção e motivação. *Revista Brasileira de Informática na Educação*, v. 13, n. 1, p. 20–34, 2005.

(Russell e Norvig 2003) RUSSELL, S.; NORVIG, P. *Artificial Intelligence: A Modern Approach*. 2. ed. New Jersey: Prentice Hall, 2003.

(Saeb, Weber e Triesch 2009) SAEB, S.; WEBER, C.; TRIESCH, J. Goal-directed learning of features and forward models. *Neural Networks*, v. 22, p. 586–592, 2009.

(Sarrafzadeh et al. 2008) SARRAFZADEH, A. et al. "How do you know that I don't understand?" A look at the future of intelligent tutoring systems. *Computers in Human Behavior*, v. 24, p. 1342–1363, 2008.

(Schank 1994) SCHANK, R. C. Active learning through multimedia. *Multimedia, IEEE*, v. 1, n. 1, p. 69–78, 1994.

(Schultz e Schultz 2005) SCHULTZ, D. P. e SCHULTZ, S. E. *História da Psicologia Moderna*. São Paulo: Pioneira Thomson Learning, 2005.

(Skinner 1950) SKINNER, B. F. Are theories of learning necessary? *Psychological Review*, v. 57, p. 193–216, 1950.

(Skinner 1969) SKINNER, B. F. *Contingencies of Reinforcement: A Theoretical Analysis*. New York: Appleton-Century-Crofts, 1969.

(Skinner 1972) SKINNER, B. F. *Tecnologia do Ensino*. São Paulo: Ed. da Universidade de São Paulo, 1972.

(Skinner 2003) SKINNER, B. F. *Ciência e Comportamento Humano*. 11. ed. São Paulo: Martins Fontes, 2003.

(Skinner 2006) SKINNER, B. F. *Sobre o Behaviorismo*. 10. ed. São Paulo: Editora Cultrix, 2006.

(Subrahmanyam et al. 2001) SUBRAHMANYAM, K. et al. The impact of computer use on children's and adolescents' development. *Applied Developmental Psychology*, v. 22, p. 7–30, 2001.

(Sutton e Barto 1998) SUTTON, R. S.; BARTO, A. G. *Reinforcement Learning*. Cambridge, Massachusetts and London, England: Bradford Book/MIT Press, 1998.

(Thierauf 1995) THIERAUF, R. J. *Virtual reality systems for business*. USA: Quorum Books Westport, 1995.

(Watkins e Dayan 1992) WATKINS, C. J. C. H.; DAYAN, P. Q-learning. *Machine Learning*, Kluwer Academic Publishers, Boston, v. 8, p. 279–292, 1992.

(Welsh, Bernstein e Luthans 1993) WELSH, D. H. B.; BERNSTEIN, D. J.; LUTHANS, F. Application of the premack principle of reinforcement to the quality performance of service employees. *Journal of Organizational Behavior Management*, v. 13, n. 1, p. 9–32, March 1993.

(Wilson e Cole 1992) WILSON, B.; COLE, P. A critical review of elaboration theory. *Educational Technology Research and Development*, v. 40, n. 3, p. 63–79, 1992. Disponível em: <http://www.cudenver.edu/ bwilson>.